Large Language Models Are Human-Like Internally

Tatsuki Kuribayashi[1,4] **Yohei Oseki**[2] **Souhaib Ben Taieb**[1,3]
Kentaro Inui[1,4,5] **Timothy Baldwin**[1,6]
[1]MBZUAI [2]The University of Tokyo [3]University of Mons
[4]Tohoku University [5]RIKEN [6]The University of Melbourne
{tatsuki.kuribayashi,souhaib.bentaieb,
kentaro.inui,timothy.baldwin}@mbzuai.ac.ae
oseki@g.ecc.u-tokyo.ac.jp

Abstract

Recent cognitive modeling studies have reported that larger language models (LMs) exhibit a poorer fit to human reading behavior (Oh and Schuler, 2023b; Shain et al., 2024; Kuribayashi et al., 2024), leading to claims of their cognitive implausibility. In this paper, we revisit this argument through the lens of mechanistic interpretability and argue that prior conclusions were skewed by an exclusive focus on the final layers of LMs. Our analysis reveals that next-word probabilities derived from internal layers of larger LMs align with human sentence processing data as well as, or better than, those from smaller LMs. This alignment holds consistently across behavioral (self-paced reading times, gaze durations, MAZE task processing times) and neurophysiological (N400 brain potentials) measures, challenging earlier mixed results and suggesting that the cognitive plausibility of larger LMs has been underestimated. Furthermore, we first identify an intriguing relationship between LM layers and human measures: earlier layers correspond more closely with fast gaze durations, while later layers better align with relatively slower signals such as N400 potentials and MAZE processing times. Our work opens new avenues for interdisciplinary research at the intersection of mechanistic interpretability and cognitive modeling.[1]

1 Introduction

Understanding human sentence processing has long been a fundamental goal in linguistics. This goal is typically approached by investigating *what computational models can simulate human sentence processing data*, such as eye movement patterns during reading, in the field of computational psycholinguistics (Crocker, 2007; Beinborn and Hollenstein, 2024). Natural language processing (NLP) models, such as neural language models (LMs), have played a crucial role in this endeavor, serving as tools to test linguistic hypotheses. Specifically, the theory of expectation-based human sentence processing (Hale, 2001; Levy, 2008; Smith and Levy, 2013) — which posits that humans continuously predict upcoming linguistic information during reading — naturally raises the following questions: how well do word probabilities (i.e., surprisal, $-\log p(\text{word}|\text{context})$) derived from LMs align with human sentence processing behavior? What kind of LMs produce the most human-like surprisal?

Previous studies have provided substantial evidence supporting expectation-based accounts of human sentence processing (Shain et al. 2024; *inter alia*). However, they reveal an intriguing trend: surprisal estimates from large language models (LLMs) often deviate from human reading behavior, and rather smaller models, such as GPT-2 small, offer better simulations of human behavior (Shain et al., 2024; Oh and Schuler, 2023b; Kuribayashi et al., 2022, 2024). This observation — *larger LMs are less human-like* — has sparked intriguing linguistic questions (Wilcox et al., 2024) as well as a fair amount of confusion within the community. Why do smaller LMs appear more human-like, despite their generally poorer linguistic competence (Waldis et al., 2024)?

In this work, we highlight the cognitively *plausible* aspects of LLMs, challenging existing conclusions. Specifically, we show that **surprisal derived from the internal layers of larger LMs aligns with human sentence processing data as well as, or even better than, that from smaller LMs**. Previous studies, focusing exclusively on final-layers' surprisal, have overlooked this critical insight. These results could be drawn

[1]Code is available at https://github.com/kuribayashi4/surprisal_internal_layers

with techniques from mechanistic interpretability (Dar et al., 2023; Belrose et al., 2023; Wendler et al., 2024), *logit lens* (nostalgebraist, 2020) or dubbed *early exits* (Kaya et al., 2019); we compute next-word surprisals directly from internal layers of LMs by projecting intermediate representations into the output vocabulary space, bypassing subsequent layers. We additionally reveal that surprisal from earlier layers fits better with fast human responses (first-pass gaze durations and self-paced reading time), while surprisal from later layers aligns more closely with slower measures (N400 and MAZE task data) (Figure 1). This also resolves the previously suggested *behavior–neurophysiology gap* in LM-based cognitive modeling: smaller LMs predict reading behavior better (Oh and Schuler, 2023b), while larger LMs excel in modeling neurophysiological data (Schrimpf et al., 2021; Michaelov et al., 2024a; Hosseini et al., 2024). We suggest that this gap stems from the inconsistent treatment of internal layers (e.g., exclusive reliance on final layers, or inconsistent inclusion of intermediate layers). If all the internal layers are focused on, even larger LMs have human-like responses through the lens of both behavioral and neurophysiological data.

Our exploration aligns with the common view that different human measures, operating on distinct timescales, reflect somewhat different stages of sentence processing. For example, fast responses, such as first-pass gaze durations (∼200ms), capture early-stage lexical processing (Calvo and Meseguer, 2002), while slower responses, such as N400 event-related potentials (∼400ms), correspond to deeper semantic integration (Lau et al., 2008; Kutas and Federmeier, 2011; Nour Eddine et al., 2024). Analogously, internal layers of LLMs may encode these temporal distinctions: earlier layers align with fast, shallow processes, while later layers correspond to slower, richer processes (Tenney et al., 2019).

In summary, our results suggest that larger LMs provide superior cognitive plausibility in modeling both human behavior and neurophysiology data internally. In other words, shallower, cognitively plausible LMs are "nested" within LLMs. Broadly, these findings advocate the integration of cognitive modeling and mechanistic interpretability, encouraging a focus on layer-wise alignment with human measures.

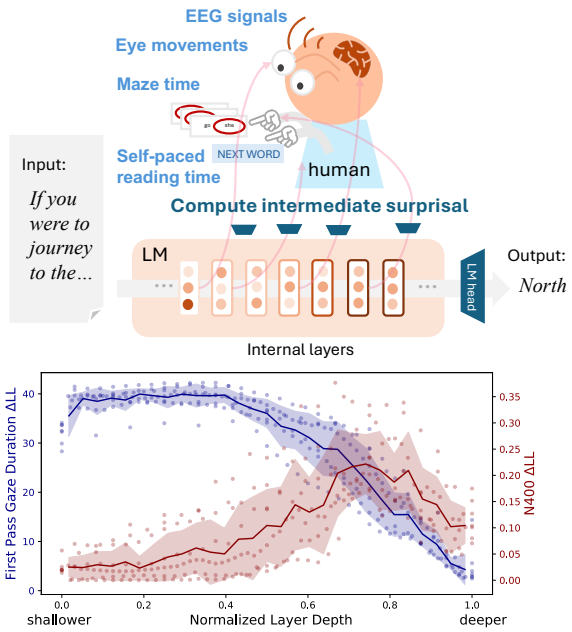


Figure 1: Different measures of human sentence processing align with surprisal from different layers of language models (LMs), and the best layer is typically not the final layer. In the bottom plot, for example, gaze duration (blue dots) and EEG signal (red dots) correlate with earlier and later layers of LMs, respectively. Each dot corresponds to the fit of surprisal (y-axis) from a particular layer depth (x-axis) to human data (ZuCO corpus).

2 Related work

2.1 Cognitive modeling and NLP

A key objective in linguistics is to understand how humans process language (Crocker, 2007), a goal that remains pertinent even in the era of LLMs. According to perspectives outlined by Marr (1982), information processing can be examined at three levels: (i) the computational level: *what is the goal of computation?*; (ii) the algorithmic level: *how does the model achieve the goal?*; and (iii) the implementational level: *how is it physically implemented?*. Humans can be viewed as an information processing model, and surprisal theory — humans continuously predict upcoming information during reading, with cognitive load incurred by unpredictable information — (Hale, 2001; Levy, 2008; Smith and Levy, 2013) has accumulated its empirical evidence (Smith and Levy 2013; Frank et al. 2015; Shain et al. 2024; *inter alia*).[2] An orthogonal, algorithmic-level ques-

[2]Surprisal theory has also been critiqued (van Schijndel and Linzen, 2021; Huang et al., 2024), particularly for its fail-

tion regarding the surprisal theory is *with what kind of algorithms and representations, humans predict upcoming information*. The NLP community has developed various methods to compute next-word probabilities, ranging from incremental parsers to LLMs, and researchers have tested their alignment with human reading data to provide insights into that question. This alignment is typically investigated by analyzing which LMs compute surprisal $-\log p(\text{word}|\text{context})$ that correlates with human measures (e.g., word-by-word gaze durations) based on the surprisal theory.

2.2 Poorer fit of larger LMs' surprisal to human reading behavior

In the days of much smaller LMs/parsers than modern LLMs (Hale 2001; Levy 2008; Smith and Levy 2013; Frank and Bod 2011; Aurnhammer and Frank 2019; Merkx and Frank 2021; *inter alia.*), model-scaling generally improved their ability to simulate human sentence processing data (Frank and Bod, 2011; Goodkind and Bicknell, 2018; Wilcox et al., 2020, 2023a). However, recent studies have questioned the generality of this scaling effect. The reversed trend was first found in typologically distant languages (Kuribayashi et al., 2021), and even within English, further scaling up LMs have shown weaker alignment with human reading behavior (Kuribayashi et al., 2022; Shain et al., 2024; Oh and Schuler, 2023b). This *bigger is not always better* phenomenon has become a key focus of LM-based cognitive modeling (Wilcox et al., 2024), with researchers investigating why LLMs appear cognitively implausible (Kuribayashi et al., 2022; Oh and Schuler, 2023a,b; Oh et al., 2024; Nair and Resnik, 2023; Kuribayashi et al., 2024). In addition, from a more interdisciplinary view, mixed results are reported regarding such scaling effects. For example, smaller LMs simulate reading behavior better (Oh and Schuler, 2023b), while larger LMs simulate neurophysiological data better (Schrimpf et al., 2021; Michaelov et al., 2024a; Hosseini et al., 2024). Opposite effect of instruction-tuning was also observed between brain data and behavioral data (Aw et al., 2024; Kuribayashi et al., 2024). We offer a perspective to address these negative scaling effects and behavior–neurophysiology gap, showing that the internal layers of larger LMs are more effective at modeling both behavioral and neurophysiological data.

ure to account for the cognitive load incurred in complex sentences. Our contribution is orthogonal to such criticism (see the Limitations section).

2.3 Human measure differences

We analyze the next-word predictions from the internal layers of LLMs in comparison with human sentence processing data. One motivation for this analysis is that different human measures, particularly at different time scales, may emphasize different stages of sentence processing (Witzel et al., 2012; Lewis and Vasishth, 2005; Vani et al., 2021; Caucheteux et al., 2023; McCurdy and Hahn, 2024). LM internal layers, which are also computed sequentially, would be a natural counterpart to such multiple stages of processing (Tenney et al., 2019). For example, eye movements reach the next word (or further) typically in ∼200ms before N400 brain signals peak at ∼400ms (Dimigen et al., 2011), suggesting that fast gaze durations may not reflect the cognitive load indexed by N400 signals (Rayner and Clifton, 2009).

3 Methods

3.1 Probabilities from internal layers

Our main proposal is to use the next-word probability $p(w_t|\boldsymbol{w}_{<t})$ of word w_t in its context $\boldsymbol{w}_{<t} = [w_1, \cdots, w_{t-1}]^\top$ from internal layers in cognitive modeling, so we begin with how to extract internal surprisal. We use two methods of *logit-lens* (nostalgebraist, 2020) and its sophisticated version of *tuned-lens* (Belrose et al., 2023). The first method, *logit-lens*, extracts the probability of a word w_t from a d-dimentaional internal representation $\boldsymbol{h}_{l,t} \in \mathbb{R}^d$ at the l-th layer and time step t, as follows:

$$\begin{aligned} p(w_t|\boldsymbol{w}_{<t}; \boldsymbol{h}_{l,t}) &= \text{LogitLens}(\boldsymbol{h}_{l,t})[\text{id}(w_t)] \\ &= \text{softmax}(\boldsymbol{W}_U \text{LayerNorm}(\boldsymbol{h}_{l,t}))[\text{id}(w_t)], \end{aligned} \quad (1)$$

where $\boldsymbol{W}_U \in \mathbb{R}^{|\mathcal{V}| \times d}$ is an unembedding matrix obtained from LM's output layer, and $|\mathcal{V}| \in \mathbb{R}$ is model's vocabulary size. Simply put, the internal representaion $\boldsymbol{h}_{l,t}$ is mapped into output vocabulary space by applying $\boldsymbol{W}_U$ (i.e., skipping subsequent layers: $\boldsymbol{h}_{l+1,t}, \cdots, \boldsymbol{h}_{\text{last},t}$), and next-word probability is obtained in that space. $\text{LayerNorm}(\cdot) : \mathbb{R}^d \rightarrow \mathbb{R}^d$ in Eq. 1 is the layer normalization at the last layer, and $[\text{id}(w_t)]$ extracts the probability for w_t[3] from the probability distribution over $\mathcal{V}$, obtained through the

[3]If a word is split into multiple subwords, accumulated surprisal is used. See Eq.2 in Kuribayashi et al. (2021).

$\mathrm{softmax}(\cdot) : \mathbb{R}^{|\mathcal{V}|} \to [0,1]^{|\mathcal{V}|}$ function. The obtained probability (Eq. 1) is converted to surprisal $-\log p(w_t|\boldsymbol{w}_{<t}; \boldsymbol{h}_{l,t})$, and then used in the regression model to predict human reading data (Section 3.2).

The second method, tuned-lens, extends logit-lens to handle the potential *representational drifts* through layers. This technique introduces an additional linear transformation for each layer l to mitigate the mismatch between the representation spaces of the l-th layer and the last layer:

$$p(w|\boldsymbol{w}_{<t}; \boldsymbol{h}_{l,t}) = \mathrm{LogitLens}(\boldsymbol{W}_l \boldsymbol{h}_{l,t} + \boldsymbol{b}_l)[\mathrm{id}(w)], \quad (2)$$

where $\boldsymbol{W}_l \in \mathbb{R}^{d \times d}$ and $\boldsymbol{b}_l \in \mathbb{R}^d$ are additionally trained to align the output of logit-lens with the last layer's next-word probability distribution on additional LM pretraining data. We use publicly available tuned-lens parameters. Notably, we do not fine-tune any part of the LMs for human data; instead, we observe the emerging correlations between next-word probabilities and human reading measures.

3.2 Psychometric predictive power

We evaluate the ability of surprisal values to predict word-by-word human cognitive responses, such as reading times or physiological signals (Section 4.1). Following prior work (Wilcox et al. 2023b; Pimentel et al. 2023, *inter alia*), this is done using linear regression models, motivated by surprisal theory (Smith and Levy, 2013; Shain et al., 2024), which posits a linear relationship between surprisal and processing cost.

Formally, let $\boldsymbol{w} = [w_1, \ldots, w_n]^\top$ denote the sequence of words in a dataset, and let $\boldsymbol{y} = [y_1, \ldots, y_n]^\top \in \mathbb{R}^n$ be the corresponding word-by-word human measurements. We assess the predictive contribution of surprisal values $\boldsymbol{s} = [s(w_1), \ldots, s(w_n)]^\top \in \mathbb{R}^n_{\geq 0}$, where surprisal is defined as $s(w_t) := -\log p(w_t|\boldsymbol{w}_{<t})$, using model probabilities as described in Section 3.1.

To evaluate the added benefit of surprisal over more primitive linguistic factors, we include a baseline feature vector $\boldsymbol{b}(w_t)$ for each word w_t:[4]

$$\boldsymbol{b}(w_t) = [\mathrm{length}(w_t), \mathrm{freq}(w_t), \mathrm{length}(w_{t-1}), \mathrm{freq}(w_{t-1}), \mathrm{length}(w_{t-2}), \mathrm{freq}(w_{t-2}), s(w_{t-1}), s(w_{t-2})]^\top. \quad (3)$$

We include features of the two preceding words to account for spillover effects — i.e., the processing difficulty of w_{t-1} or w_{t-2} can influence the response to w_t.

We train two nested linear regression models[5]: (i) a full model including both surprisal and baseline features; and (ii) a reduced model using only the baseline features. The coefficients are estimated via ordinary least squares. Model fit is quantified by the log-likelihood under a Gaussian noise assumption, and the difference in log-likelihoods between the two models — denoted $\Delta\mathrm{LL}$ — reflects the isolated contribution of surprisal.

A higher $\Delta\mathrm{LL}$ indicates greater predictive power, and we refer to this value as the *psychometric predictive power* (PPP). The central question of our study is: which LM layer yields surprisal values with the strongest PPP?

4 Experimental settings

4.1 Human data

We use 15 human reading datasets, listed in Table 1, (we additionally use MECO in Section 6.5), which include human measurements from various methods: self-paced reading time (SPR), first-pass gaze duration (FPGD), Maze task processing time (MAZE), and electroencephalography (EEG; specifically the N400 component). The datasets share a common format: each word w_t is annotated with $\mathrm{Cost}(w_t) \in \mathbb{R}$ representing the human cognitive load associated with it. Our corpus selection aligns with recent studies (Kuribayashi et al., 2024; Michaelov et al., 2024a; de Varda et al., 2024; McCurdy and Hahn, 2024).[6]

[4] Word length is measured in characters; word frequency is estimated using `word_freq` (Speer, 2022). We use a consistent set of baseline features across all datasets and models, with a few exceptions. For N400 data, a baseline amplitude term is added; for Michaelov et al. (2024b)'s EEG data, we additionally include electrode-level random effects.

[5] Implemented using the `statsmodels` package (Seabold and Perktold, 2010). Some existing works include preceding words' surprisal values $s(w_{t-1}), s(w_{t-2})$ only in the full regression model (not in the reduced one; Eq. 3) when computing $\Delta\mathrm{LL}$. We confirmed that this setting variation does not alter the conclusion, and at least in this paper, we adopt the setting to include these features in both full and reduced regression models.

[6] We applied the same preprocessing as Kuribayashi et al. (2024) (DC, NS), de Varda et al. (2024) (UCL), Hahn et al. (2022) (Fillers), and Michaelov et al. (2024a) (N400). For ZuCO, we only used the naturalistic reading part, and for its N400, we averaged the values at the central electrode between 300-500ms during the first pass over a word.

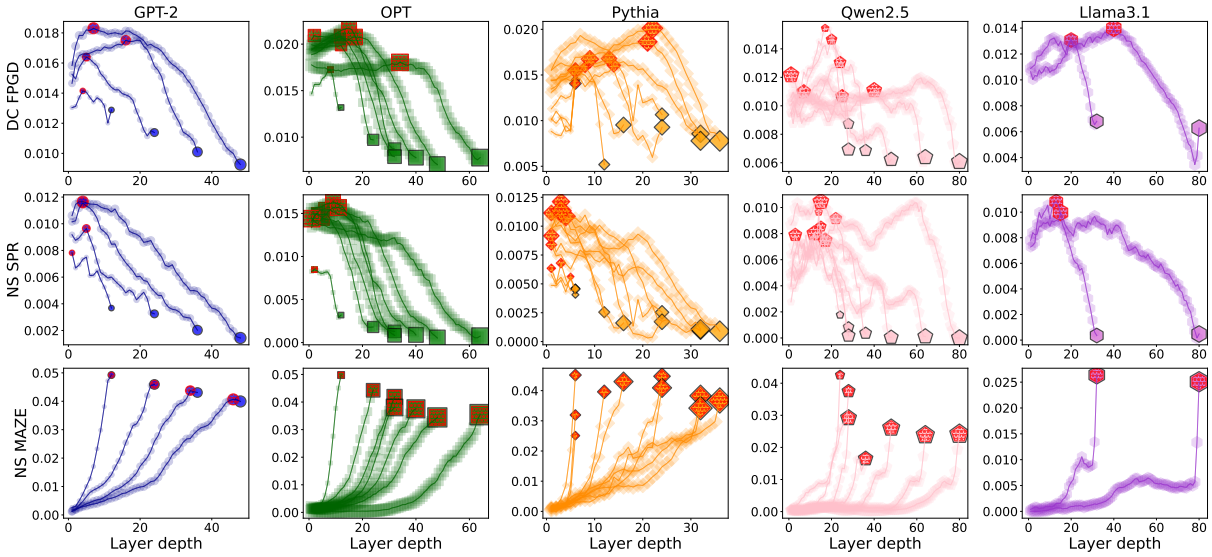


Figure 2: Relationships between layer depth (x-axis) and ΔLL (y-axis) for each LM in three datasets: FPGD in DC, SPR in NS, and MAZE in NS. The graphs are separated by model families and data. Each line corresponds to a different model, and one with larger markers and more layers corresponds to the bigger model within each model family. For each model, the best layer is indicated with a red edge line and shading patterns, and the last layer is indicated with a black edge. The graph starts at the first layer, not at the embedding layer. The results are based on logit-lens.

SPR is measured by presenting sentences through a sliding word-by-word window, with participants pressing a button to advance. FPGD, a key eye-tracking measure, represents the total time from first fixating on a word to moving to another word. Maze processing time is measured during a task requiring participants to select the plausible continuation of a sentence, offering a controlled alternative to naturalistic reading. EEG measures brain activity, with N400 reflecting the negative brain potential peaking around 400ms after word presentation. These are the common measures employed to study expectation-based sentence processing. SPR, FPGD, and MAZE are categorized as human *behavioral* data, while EEG falls under *neurophysiological* data.

To minimize confounding factors between stimulus data and human measures, we included datasets with multi-layered annotations across multiple human measures. These include the Natural Stories Corpus (Futrell et al., 2021) with SPR[7] and MAZE data (Boyce and Levy, 2023), ZuCO corpus (Hollenstein et al., 2018) with FPGD and N400 data, UCL Corpus (Frank et al., 2013) annotated with SPR, FPGD, and N400 data (Frank et al., 2015), and filler sentences from Hahn et al. (2022) annotated with SPR, FPGD, and MAZE data (Vasishth et al., 2010; Hahn et al., 2022). In particular, the FPGD and N400 data in ZuCO were simultaneously recorded from the same human subjects, which likely minimized confounding factors. As is common in preprocessing, we exclude data points with zero SPR/FPGD/MAZE value. Human data for each token in the corpus were averaged across subjects prior to analysis, following recent practices (Pimentel et al., 2023; Oh and Schuler, 2023b; Kuribayashi et al., 2024; de Varda et al., 2024).

[7]We use the version (2025-05-12) without the misalignment problem (see https://github.com/languageMIT/naturalstories).

4.2 Language models

We evaluate 30 open-source LMs including bilion-scale ones: 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 (14M, 31M, 70M, 160M, 410M, 1B, 1.4B, 2.8B, 6.9B, and 12B parameters) (Biderman et al., 2023), Qwen2.5 (0.5B, 1.5B, 3B, 7B, 14B, 32B, 72B), and Llama-3.1 (8B and 70B). See Appendix B for details. For tuned-lens experiments (Section 3.1), we use 14 of these models based on the availability of pre-

Stimuli	Measure	PPP_logit-lens					PPP_tuned-lens				
		0-0.2	0.2-0.4	0.4-0.6	0.6-0.8	0.8-1	0-0.2	0.2-0.4	0.4-0.6	0.6-0.8	0.8-1
DC	FPGD (Kennedy et al., 2003)	14.13	14.78	**14.92**	13.38	9.84	**17.10**	16.32	15.39	13.53	10.49
NS	SPR (Futrell et al., 2021)	**9.85**	9.75	8.44	5.68	2.67	**8.93**	7.03	5.11	3.44	2.33
	MAZE (Boyce and Levy, 2023)	1.18	3.00	5.69	12.06	**23.77**	9.70	17.56	24.15	32.86	**39.63**
ZuCO	FPGD (Hollenstein et al., 2018)	38.10	**38.13**	35.59	29.82	15.94	**30.48**	27.16	22.56	17.29	8.77
	N400 (Hollenstein et al., 2018)	0.07	0.12	0.15	**0.18**	0.16	0.20	0.32	**0.34**	0.29	0.18
UCL	SPR (Frank et al., 2013)	**22.88**	22.21	19.30	11.45	4.77	**15.78**	8.92	4.87	2.53	1.27
	FPGD (Frank et al., 2013)	22.11	**23.39**	22.83	15.77	6.53	**16.28**	14.48	11.87	9.47	5.57
	N400 (Frank et al., 2015)	**56.86**	38.04	22.77	16.07	22.58	11.31	6.12	16.19	29.49	**37.11**
Fillers	SPR (Vasishth et al., 2010)	7.83	10.89	**14.39**	14.18	14.35	8.60	10.47	11.36	11.86	**13.33**
	FPGD (Vasishth et al., 2010)	6.66	5.83	6.48	7.31	**10.36**	8.94	10.91	12.91	13.81	**14.00**
	MAZE (Hahn et al., 2022)	5.39	3.01	5.08	21.97	**60.89**	9.96	28.27	52.00	73.38	**88.64**
Michaelov+,2024	N400 (Michaelov et al., 2024b)	0.88	1.42	**1.91**	1.68	0.91	0.95	1.51	**1.70**	1.38	0.99
Federmeier+,2007	N400 (Federmeier et al., 2007)	0.77	3.11	8.59	18.05	**25.80**	1.49	5.22	13.06	24.48	**28.71**
W&F,2012	N400 (Wlotko and Federmeier, 2012)	**0.35**	0.19	0.10	0.09	0.12	**0.51**	0.27	0.12	0.05	0.11
Hubbard+,2019	N400 (Hubbard et al., 2019)	0.18	0.23	0.23	**0.25**	0.17	0.11	0.12	0.22	**0.36**	0.33
S&F,2022	N400 (Szewczyk and Federmeier, 2022)	0.11	0.15	0.38	0.90	**1.40**	0.16	0.33	0.77	1.29	**1.42**
Szewczyk+,2022	N400 (Szewczyk et al., 2022)	1.21	2.91	4.43	6.40	**8.04**	2.12	3.58	5.52	8.10	**8.93**

Table 1: All the results. The ΔLL scores are averaged by the layer relative depth, e.g., first 20% of layers as "0-0.2," across models, and the best relative layer range for each data is highlighted in bold. ΔLLs are multiplied by 1000 for brevity.

trained tuned lenses.[8] The number of internal layers in our models ranges from 6 to 80. Results for surprisal from the embedding layer are excluded as these generally show a bad fit with human data.

5 Experimental results

Figures 2 (Section 5.1) and 3 (Section 5.2) summarize our main findings, with comprehensive results in Table 1.

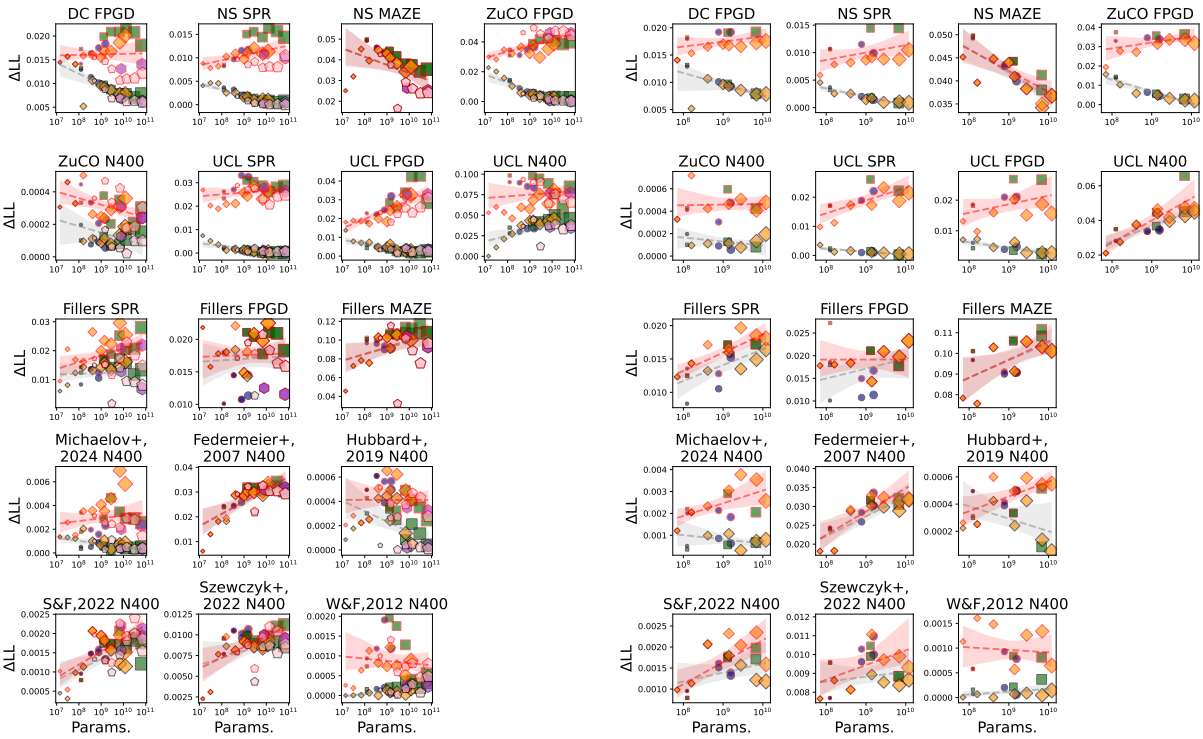
5.1 Last layer does not yield the best ΔLL

We first revisit the experimental settings (DC and NS datasets) of Oh and Schuler (2023b). The top two line graphs in Figure 2 depict the layer-wise ΔLL for each LM. A consistent pattern emerges in FPGD and SPR data: the ΔLL decreases toward the final layer (the rightmost, less-transparent, large markers), indicating that the last layer often yields the lowest score compared to the internal layers of the same model. These results challenge the assumption, widely adopted in existing studies, that the last layer is the most reliable indicator of an LM's cognitive plausibility.

At the same time, changing the human measurement, i.e., from SPR to MAZE in the NS data (the second and third rows in Figure 2), flips the trend. This indicates that the aligned layers can easily change when using different human measurement methods, and suggests that different aspects of human sentence processing correlate with different LM layers, depending on the human measurement method. Such measurement–layer interaction effect is further looked into in Section 6.2.

Table 1 presents a detailed breakdown of all the datasets, averaging ΔLL scores across relative layer positions (e.g., 0–0.2 for the first 20% of layers). Table 1 includes the results with logit-lens and tuned-lens, which yielded generally consistent patterns. As shown in Figure 2, the best-performing layer is often not the final one (0.8–1.0), and the optimal layer varies based on the type of measurement and stimuli. For instance, SPR and FPGD data are best modeled by earlier layers, whereas MAZE processing times and N400 signals are better captured by later layers (as motivated in Section 2.3). Furthermore, there are also stimulus-dependent biases — the optimal layer for the UCL dataset is among the earlier layers, while for the Fillers dataset, it lies among the later layers (perhaps associated with the specific complexity

[8] We used parameters released in https://huggingface.co/spaces/AlignmentResearch/tuned-lens/tree/main. Specifically, we used GPT-2 124M, 774M, 1.5B; OPT 125M, 1.3B, 6.7B; and Pythia 70M, 160M, 410M, 1B, 1.4B, 2.8B, 6.9B, 12B, in the tuned-lens experiments. Notably, their actual implementation used the representation of $\boldsymbol{h}_{l,t}\boldsymbol{W}'_l + \boldsymbol{h}_{l,t} + \boldsymbol{b}_l = \boldsymbol{h}_{l,t}(\boldsymbol{W}'_l + \mathbf{1}) + \boldsymbol{b}_l$, but we omit the identity matrix $\mathbf{1} \in \mathbb{R}^{d \times d}$ in Eq. 2 by overriding $\boldsymbol{W}_l = \boldsymbol{W}'_l + \mathbf{1}$.



(a) Logit-lens

(b) Tuned-lens

Figure 3: Scaling effect between model size (parameter counts in log scale) and ΔLL. Each marker corresponds to each LM's ΔLL score from its best layer (red edge) or the last layer (black edge). The regression lines show the scaling effects, and the red line is for best-layer's ΔLL while the grey one is for the last layer's. The maker type (shape/size/color) follows Figure 1.

of the stimulus).

5.2 Revisiting LM-scaling effects in cognitive modeling with internal layers

We revisit the question with our extended focus on model internals: what kind of LMs yield the best ΔLL from their internals? As the field is particularly interested in the relationship with model scaling (Goodkind and Bicknell, 2018; Oh and Schuler, 2023b), we examine the relationship between LM parameter size (x-axis) and ΔLL (y-axis) for two scenarios: (1) using the last layer's ΔLL (grey lines), reproducing previous findings; and (2) considering the best ΔLL layer identified in this study (red lines).

Figure 3 illustrates these two relationships. The grey lines align with prior findings relying on the last layer (Oh and Schuler, 2023b; Michaelov et al., 2024a), showing mixed scaling effects, where larger LMs do not consistently outperform smaller ones. However, the red lines reveal a positive scaling trend: larger LMs achieve equal or better ΔLL compared to smaller LMs when internal layers are taken into account. The Pearson correlation coefficients between parameter numbers and ΔLL from the best layers were significantly larger than zero on average, across settings.[9] This suggests that when the analysis extends to internal layers, the ΔLL ranking flips, revealing that larger LMs are seemingly more cognitively plausible. In other words, larger LMs embed cognitively plausible, smaller LMs within their internal. One notable exception is the MAZE processing time in the NS dataset (Boyce and Levy, 2023), where a strictly negative scaling effect persists, even when internal layers are considered. PPL–ΔLL relationships[10] are additionally shown in Figure 8 in the Appendix, which also show that the poor ΔLL of

[9] We collected the correlation coefficients between logarithmic number of model parameters and ΔLL from the best layers from 34 settings of $\{\text{dataset}\} \times \{\text{lens}\}$, and one-sample t-test shows that these coefficients are, on average, significantly larger than zero (p-value < 0.05).

[10] Perplexity (PPL), a general quality measure of LMs, is a geometric mean of next-word probabilities over data L: $\prod_{t=1}^{|L|} p(w_t|\boldsymbol{w}_{<t})^{1/|L|}$. The PPL–ΔLL relationship has long been investigated (Frank and Bod, 2011; Goodkind and Bicknell, 2018; Kuribayashi et al., 2021; Oh and Schuler, 2023b).

larger, more accurate LMs is mitigated.

6 Analyses

We conduct several follow-up analyses to support and clarify the findings reported in Section 5.

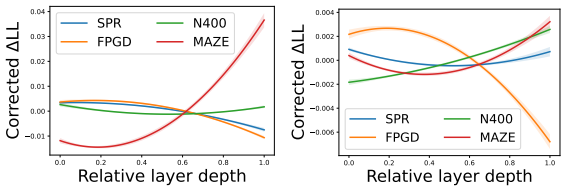
6.1 How easily can good layers be found?

One immediate concern in Section 5.2 is how many numbers of internal layers yield a good ΔLL; perhaps, we just observed outliers as best-ΔLL values in Figure 3. Given this concern, we analyze the number of internal layers that outperform the previously best ΔLL achieved by the last layer within the same model family. Table 2 presents the win rate of internal layers' ΔLL against the respective previous best score. The win rate is typically around 80%, indicating a significant proportion of internal layers achieved good ΔLL scores. These findings support that the cognitive plausibility of LLMs has been underestimated and that our argument (Figure 3) was not based on specific outlier layers but reflected a broader trend across many internal layers.

6.2 Layer depth and human measures

We observed systematic tendencies in the relationship between layer depth and human measurement methods. For instance, FPGD aligns better with earlier layers, whereas N400 aligns better with later layers, as summarized in Table 1. To statistically validate this relationship, we trained a linear regression model to explain ΔLL scores from our 23,154 experimental settings $\{\text{dataset}\} \times \{\text{model}\} \times \{\text{layer}\}$ with the following features for each setting s: $\{\text{stimuli}(s)$, $\text{model}(s)$, $\text{lens}(s)$, $\text{layer\_depth}(s)$, $\text{measure}(s)$, $\text{layer\_depth}(s)$:$\text{measure}(s)\}$. Here, stimuli represents the source stimuli of the data ("Stimuli" column in Table 1), model encodes the model name, layer_depth is the depth of the layer where the ΔLL is obtained, measure encodes the human measurement method ("Measure" column in Table 1), and lens indicates whether the logit-lens or tuned-lens was used. The term layer_depth $\times$ measure captures the interaction between effective layer depth and human measures, which is of interest. Note that measure is a categorical variable, with SPR serving as the dummy category.

The coefficients for layer_depth $\times$ N400 and layer_depth $\times$ MAZE were significantly larger than zero (p-value $<$ 0.001), while that for layer_depth $\times$ FPGD is significantly lower than layer_depth $\times$ SPR (see full regression results in Table 5 in the Appendix). This confirms that SPR aligns with earlier layers than those yielding a good fit with FPGD, and N400 and MAZE align with later layers than FPGD. We also visualize the relationships between ΔLL and relative layer depth for each human measure in Figure 4a (polynomial fit using 2nd-order term). Here, we use corrected ΔLLs that are computed by subtracting variances explained by factors other than measure based on the regression model. The lines also indicate the differences across different human measures, e.g., good ΔLL for MAZE is clearly associated with the latter layers.



(a) All tokens. (b) Clause-final tokens.

Figure 4: Relationship between ΔLL and relative layer depth for each human measure. Different measures are associated with different layers; for example, good ΔLLs for FPGD are achieved in earlier layers, while those for MAZE are in the latter layers.

6.3 Are results biased by targeted tokens?

A potential confound in Section 6.2 stems from differences in targeted tokens for different human measures. For instance, N400 data are typically recorded only at sentence-final tokens to preprocess continuous, time-series EEG data, potentially introducing biases specific to sentence-final tokens, such as wrap-up effects (Just and Carpenter, 1980; Rayner et al., 2000; Meister et al., 2022). In contrast, behavioral measures are recorded across tokens within a sentence. To address this potential confound, we conducted additional experiments targeting sentence/clause-final tokens for all measures, even for SPR, FPGD, and MAZE. Sentence/clause-final tokens are ideitified using a constituency parser (Kitaev et al., 2019; Kitaev and Klein, 2018) and punctuations (e.g., ".", ","), following Meister et al. (2022).

Table 3 presents the results. Even when analyses were restricted to sentence-final tokens, earlier layers continued to align better with SPR

	GPT2	OPT						Pythia					Qwen				Llama-3	
Data	XL	1.3B	2.7B	6.7B	13B	30B	66B	1B	1.4B	2.8B	6.9B	12B	3B	7B	32B	72B	8B	70B
DC FPGD (Kennedy et al., 2003)	0.80	0.80	0.82	0.76	0.73	0.76	0.78	0.00	0.32	0.36	0.73	0.81	0.10	0.26	0.72	0.94	0.89	0.73
NS SPR (Futrell et al., 2021)	0.82	0.80	0.82	0.76	0.76	0.78	0.82	0.65	0.60	0.64	0.73	0.95	0.93	0.95	0.93	0.97	0.98	0.73
ZuCO FPGD (Hollenstein et al., 2018)	0.80	0.84	0.88	0.79	0.73	0.78	0.86	0.65	0.60	0.55	0.76	0.97	0.97	0.97	0.98	0.97	0.98	0.78
UCL SPR (Frank et al., 2013)	0.78	0.80	0.79	0.76	0.76	0.78	0.80	0.71	0.52	0.64	0.70	0.97	0.97	0.97	0.98	0.97	0.99	0.70
UCL FPGD (Frank et al., 2013)	0.94	0.88	0.82	0.79	0.83	0.88	0.83	0.59	0.56	0.58	0.70	0.97	0.93	0.97	0.94	0.97	0.99	0.73

Table 2: How likely ΔLL from internal layers outperformed the previous best ΔLL (achieved within the same model family, relying on their last layers). The results are focused on billion-scale models and behavioral data with somewhat drastic flips in LM-scaling effects for ΔLLs.

		Logit-lens (ΔLL)					Tuned-lens (ΔLL)				
Stimuli	Measure	0-0.2	0.2-0.4	0.4-0.6	0.6-0.8	0.8-1.0	0-0.2	0.2-0.4	0.4-0.6	0.6-0.8	0.8-1.0
DC	FPGD (Kennedy et al., 2003)	1.24	1.37	**1.42**	1.31	1.00	1.59	1.60	**1.65**	1.58	1.30
NS	SPR (Futrell et al., 2021)	**0.45**	0.38	0.33	0.29	0.21	0.16	0.12	0.12	0.18	**0.25**
	MAZE (Boyce and Levy, 2023)	0.35	0.36	0.42	0.62	**0.84**	0.67	0.70	0.76	1.02	**1.30**
ZuCO	FPGD (Hollenstein et al., 2018)	38.10	**38.13**	35.59	29.82	15.94	**30.48**	27.16	22.56	17.29	8.77
	N400 (Hollenstein et al., 2018)	0.07	0.12	0.15	**0.18**	0.16	0.20	0.32	**0.34**	0.29	0.18
UCL	SPR (Frank et al., 2013)	**3.09**	2.53	2.12	1.21	0.50	**1.76**	0.89	0.50	0.27	0.15
	FPGD (Frank et al., 2013)	7.51	7.89	**8.33**	6.67	3.56	4.85	**4.86**	4.60	4.26	2.66
	N400 (Frank et al., 2015)	3.58	1.66	1.99	5.20	**11.95**	1.12	3.85	8.31	12.75	**18.17**
Fillers	SPR (Vasishth et al., 2010)	0.28	0.33	0.62	1.86	**5.78**	0.88	2.19	3.32	5.83	**10.31**
	FPGD (Vasishth et al., 2010)	0.30	0.25	0.26	0.59	**1.29**	0.28	0.42	0.52	1.40	**2.04**
	MAZE (Hahn et al., 2022)	3.54	2.41	1.51	3.83	**8.99**	1.15	1.94	5.52	10.64	**12.53**

Table 3: Results for the same settings as Table 1, except that only sentence/clause-final tokens are targeted. Some N400 data are omitted because they initially targeted only sentence/clause-final tokens. ΔLLs are multiplied by 1000 for brevity.

and FPGD data. Regression analysis (same as Section 6.2) confirmed that the coefficients for $\text{layer\_depth} \times \text{N400}$ and $\text{layer\_depth} \times \text{MAZE}$ remained significantly larger than those for $\text{layer\_depth} \times \text{SPR}$, and the coefficient for $\text{layer\_depth} \times \text{FPGD}$ was lower than that for $\text{layer\_depth} \times \text{SPR}$ (see full regression results in Table 5b in the Appendix). These patterns are visualized in Figure 4b.

6.4 When are earlier layers advantageous?

We further explore more general trends on when and why earlier layers' suprisal aligns better with human reading data. Following Oh and Schuler (2023b), we analyze by-token squared residual errors from regression models predicting human data (Section 3.2). We specifically use the largest data of DC (FPGD) and identify tokens where the use of the best internal layer, rather than the last layer, notably reduces errors.[11] To address this question, we fit a linear regression model to explain the decreases in squared residual errors observed in each data point w_t with an LM θ, with the following by-token linguistic properties as features: $\{\text{model}(\theta), \text{length}(w_t), \text{freq}(w_t), \text{position}(w_t), \text{POS}(w_t)\}$.

Results show that decreases in modeling error are associated with less frequent, longer words (see Table 6 in Appendix for full regression results). This aligns with prior observations (Oh et al., 2024) that LLMs tend to predict infrequent tokens with overly confident surprisals, and surprisal from internal layers mitigates this issue.

6.5 Mutlilingual generallity

Our main experiments were limited to the English language — what about the advantage of internal surprisal in other languages? Following the experiments of de Varda and Marelli (2023), we also explored the multilingual generality of our re-

[11] Oh and Schuler (2023b) reported a misalignment between MSEs (residual errors) and log-likelihood scores due to the Euclidean norm penalty adopted in the `lme4` package. This did not arise in our analysis with `statsmodels`.

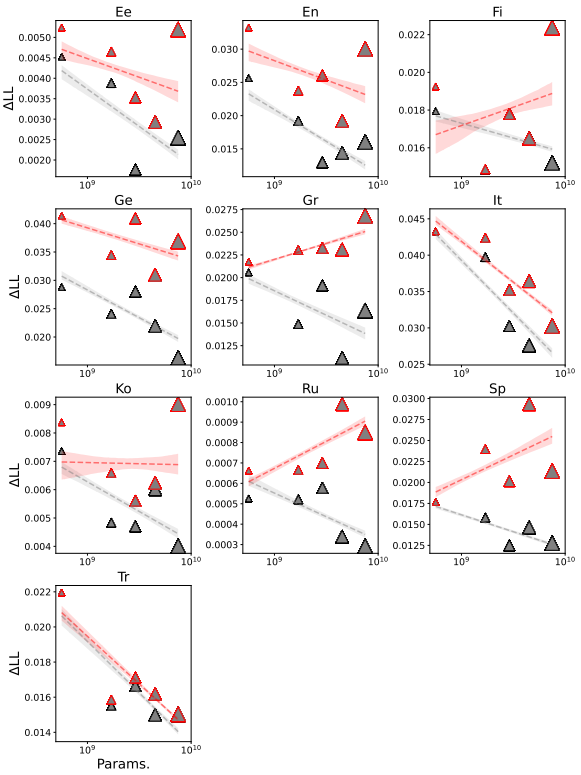
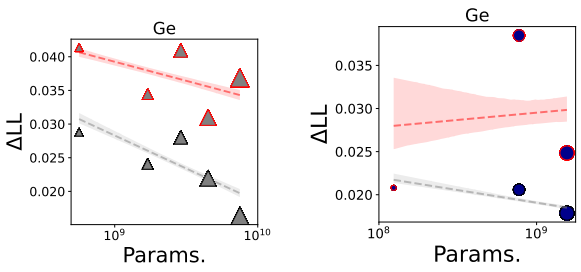


Figure 5: Scaling effect between ΔLL and parameter counts in MECO. The grey lines are results relying on the last layer's ΔLLs, and the red lines rely on the best internal layers' ΔLLs. In all the languages, the negative correlation between parameter size and ΔLL is mitigated to some extent, and in four languages, the relationship turned out to be positive.

sults, using FPGD data in ten languages recorded in MECO (Siegelman et al., 2022) Five variants of multilingual XGLMs (Lin et al., 2022) are used.[12] Figure 5 shows the relationship between parameter size and ΔLL, where the red markers and lines are based on the best internal layers, and the gray ones are from the final layers. First, the best layer typically outperforms the last layer of the respective model (86%=43/50 of the settings), as the best ΔLLs (red markers) are generally higher than the last layer's ΔLLs (grey markers) in Figure 5, reproducing the findings in Section 5.1.

Second, the negative correlation score between parameter size and ΔLL consistently increases in all the languages, which partially aligns with Section 5.2. Nevertheless, the scaling effect is still

[12]We excluded Dutch, Hebrew, and Norwegian because these are not supported by XGLMs. In addition, due to the unavailability of tuned-lens for XGLMs, only logit-lens was used.



(a) Multilingual models (b) Monolingual models

Figure 6: Scaling effects in MECO German part

negative in some languages. We suspect that these mixed results might be biased by multilingual LMs, which are reported to process every language within the English subspace in their middle layers (Wendler et al., 2024), leading to biased probability estimates. As a case study to handle this concern, we compared the results of multilingual XGMs (left part of Figure 6) with those of German monolingual GPT-2s (right part of Figure 6; see Appendix B for model details). The negative scaling effects are flipped to be positive with the use of monolingual LMs, implying that the inclusion of monolingual LMs further enhances the advantage of the internal layer's surprisal.

7 Discussion

We lastly discuss the implications of our findings, connections to existing studies, and future works.

7.1 Connection to context sensitivity

We propose a perspective linking our findings to the context sensitivity of humans and LMs during sentence processing. Earlier layers may better model human reading behavior because they are less contextualized, reflecting the human-like tendency to process sentences under working memory constraints.

Working memory limitations in humans Human sentence processing is constrained by limited cognitive resources, relying on selective and efficient use of context (Lewis and Vasishth, 2005; Lieder and Griffiths, 2019; Futrell et al., 2020; Hahn et al., 2022). Recent studies indicate that the MAZE task imposes greater working memory demands, requiring more extensive context use than SPR or FPGD (Hahn et al., 2022; McCurdy and Hahn, 2024). This aligns with our observation that MAZE processing times are better modeled by later, more context-sensitive layers (see

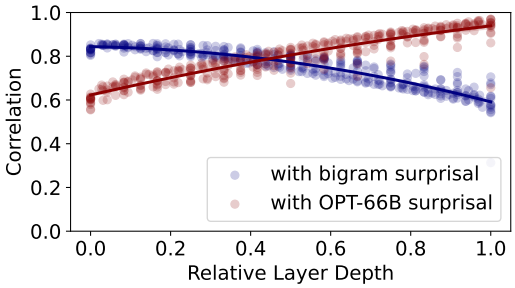


Figure 7: The markers correspond to all the internal layers of our targeted LMs, which are sorted by relative layer depth (x-axis). Two types of scores (y-axis) are plotted: (i) Pearson correlation coefficient between each layer's surprisal vs. less-contextualized bigram surprisal (blue); and (ii) each layer's surprisal vs. well-contextualized LLM surprisal (red). We used tuned-lens results.

the next paragraph), whereas SPR and FPGD align with earlier, less contextualized layers.

Working memory limitations in LMs Modern neural LMs are not optimized to conserve cognitive resources and often rely excessively on context information, resulting in superhuman predictions (Kuribayashi et al., 2022; Oh et al., 2024). However, internal layers may exhibit a human-like moderation of context use. The NLP community has observed that LMs gradually enhance contextualization across layers, from shallow representations in early layers to deeply contextualized representations in later layers (Brunner et al. 2019; Ethayarajh 2019; Toneva and Wehbe 2019).

We also confirmed the by-layer context-sensitivity of surprisal by analyzing the correlation between (i) intermediate layers' surprisal and less-contextualized bigram surprisal[13]; and (ii) intermediate layers' surprisal and more contextualized surprisal from the LLM with the lowest PPL (OPT-66B). Figure 7 shows the above two types of correlations for each model's layer, with the x-axis as the relative layer depth. This shows that earlier layers correlate more strongly with bigram surprisal (Pearson correlation coefficient r for this relationship is -0.92), while later layers align with more accurate, well-contextualized surprisal (r=0.95). These findings reinforce the idea that earlier layers exhibit limited context sensitivity, while later layers are more contextualized and better suited for modeling data like MAZE, which demands higher contextualization.

[13]Bigram LM is trained on OpenWebText (Gokaslan and Cohen, 2019) with the KenLM toolkit (Heafield, 2011)

7.2 Connection to spill-over effects

There is another possible explanation about the alignment of earlier/later layers with SPR/MAZE that was discussed in Section 6.4. Self-paced reading and eye-tracking measures often exhibit *spillover effects*, where the processing of one word influences subsequent words (Rayner, 1998). This suggests that the comprehension of a word extends beyond the immediate moment, and the associated reading times may only capture an early stage of processing. In contrast, the MAZE task (Forster et al., 2009; Boyce and Levy, 2023), which measures the time taken to select a plausible continuation from two candidates, mitigates spillover effects and is thus expected to reflect the full process of word processing. Our findings — early LM layers are more closely aligned with gaze durations and self-paced reading times, while later layers show a stronger alignment with MAZE — align with this perspective.

7.3 Connection to LM-brain alignment study

Brain imaging data (e.g., fMRI) and reading behavior data have complementary advantages. Generally speaking, the former has a high spatial resolution (in which part of the brain particular processing is performed), while the latter has a high temporal resolution (how long does the processing take). In the former context of fMRI modeling research, layer-wise LM-human alignments, similarly to us, have been attempted (Toneva and Wehbe, 2019; Schrimpf et al., 2021; Caucheteux et al., 2023) and suggested that different brain areas better align with different LM layers. Our study is orthogonal to these studies with more focus on the temporal alignment between reading-time and layer-time scales, which is concurrently explored in Hu et al. (2025).

One additional difference with the above-mentioned fMRI studies is that they typically trained linear regression models to predict brain activity directly using d-dimentional LM internal representations $\boldsymbol{h} \in \mathbb{R}^d$ as features, instead of using surprisal measures $-\log p(\text{word}|\text{context}) \in \mathbb{R}_{\geq 0}$. Thus, their results are not directly comparable with existing surprisal-based studies and may rather suffer from a confounding factor of different d for different LMs when precisely discussing LM-scaling effects.

7.4 Connection to early exits of neural models' prediction

In the engineering context, predictions from internal layers of deep neural networks (i.e., early exits of the results from internal layers) are used to improve inference efficiency by avoiding their overthinking (Graves, 2016; Banino et al., 2021), which is also called adaptive computation time (Kaya et al., 2019; Zhou et al., 2020). Such a technique has also recently been employed to enhance interpretability research to identify at which layer a particular prediction shapes (nostalgebraist, 2020; Dar et al., 2023; Belrose et al., 2023).

7.5 Limitations toward surprisal theory

Recent studies have raised several issues, orthogonal to our study, on surprisal-based cognitive modeling, for example, on tokenizations (Nair and Resnik, 2023; Giulianelli et al., 2024a; Oh and Schuler, 2024), LM training scenario (Oh and Schuler, 2023a), as well as more refined indicators of word predictability (Pimentel et al., 2023; Giulianelli et al., 2024b; Opedal et al., 2024; Meister et al., 2024). More critically, the reliance solely on surprisal obtained from LMs tends to underestimate the significant slowdown of sentence processing against syntactic amguity and grammatical violation (Wilcox et al., 2021; van Schijndel and Linzen, 2021; Huang et al., 2024; Wang et al., 2024). This study, as an initial foray, is focused on cognitive modeling on naturalistic reading corpora, but including such experiments with controlled materials will enrich our findings on internal surprisals.

Despite such concerns regarding surprisal theory, we would like to highlight that our core proposal — using probability from internal layers in cognitive modeling — is not limited to surprisal theory in human sentence processing modeling, but can contribute to any cognitive modeling work that uses probability-related measures and neural models. For example, our analysis can easily be extended to other metrics such as entropy, entropy reduction (Hale, 2016), or any new probability-based measures combining probabilities from different layers (Hu et al., 2025). Furthermore, our idea to focus on internal layers may also be combined with linguistically-motivated neural models, such as deep neural incremental parsers (Dyer et al., 2016; Yoshida et al., 2021), to potentially combine linguistic theories (e.g., theory of syntax) with surprisal theory. Thus, any limitations of surprisal theory itself do not undermine the core contributions of this work.

8 Conclusions

Recent cognitive modeling studies have demonstrated a worse fit of surprisal from larger LMs to human reading time. In this paper, we argue that these negative results stem from the exclusive focus on the *final layers* of LMs. Instead, we observe that those from *internal layers* comparably or even better fit with human behavior and neurophysiology data, suggesting that the cognitive plausibility of larger LMs has been underestimated. Furthermore, different human measurements align with different layers, implying an intriguing parallel between temporal dynamics in human sentence processing and LM internal layers.

Limitations

The experiments can be extended to more types of human measures, such as first fixation time, go-past reading time, eye regressions, total fixation time, EEG data other than the N400 potential (Federmeier et al., 2007) as well as FMRI data (Shain et al., 2022). As a first step, we have begun with the measures of SPR, FPGD, N400, and MAZE, as they are typically used in cognitive modeling. We also conducted an exploratory analysis with other human measures (e.g., P600, second-pass gaze duration) on the UCL corpus (Table 7 in Appendix). For such an extended scope, we cannot confirm the implied relationship between fast–slow human responses and early–late LM layers, and our findings appear to hold primarily for commonly-analyzed human measures of SPR, FPGD, N400, and MAZE. It is also a critical question in psycholinguistics whether the distinction between "fast" (gaze durations) and "slow" (N400) human measures adopted in this study is truly related to some cognitive costs or the time indeed humans took to process the word. Especially for EEG, the delay may perhaps be just due to technical properties regarding the temporal resolutions.

Acknowledgment

We appreciate our TACL action editor and reviewers. We also thank Ryo Yoshida for constructive feedback on the earlier version of this work. This

work was supported by JSPS KAKENHI Grant Number JP24H00087; JST PRESTO Grant Number JPMJPR21C2; JSPS Grant-in-Aid for Early-Career Scientists Grant Number JP23K16938.

References

C Aurnhammer and S L Frank. 2019. Comparing gated and simple recurrent neural network architectures as models of human sentence processing. In *Proceedings of CogSci*, pages 112–118.

Khai Loong Aw, Syrielle Montariol, Badr AlKhamissi, Martin Schrimpf, and Antoine Bosselut. 2024. Instruction-tuning aligns LLMs to the human brain. In *COLM 2024*.

Andrea Banino, Jan Balaguer, and Charles Blundell. 2021. PonderNet: Learning to ponder. In *8th ICML Workshop on Automated Machine Learning (AutoML)*.

Lisa Beinborn and Nora Hollenstein. 2024. *Cognitive plausibility in natural language processing*. Synthesis lectures on human language technologies. Springer International Publishing, Cham.

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, et al. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *ICML 2023*, pages 2397–2430. PMLR.

Veronica Boyce and Roger Philip Levy. 2023. A-maze of natural stories: Comprehension and surprisal in the maze task. *Glossa Psycholinguistics*.

Gino Brunner, Yang Liu, Damian Pascual, Oliver Richter, Massimiliano Ciaramita, and Roger Wattenhofer. 2019. On identifiability in transformers. In *ICLR 2019*.

Manuel G Calvo and Enrique Meseguer. 2002. Eye movements and processing stages in reading: Relative contribution of visual, lexical, and contextual factors. *The Spanish Journal of Psychology*, 5(1):66–77.

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

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

Guy Dar, Mor Geva, Ankit Gupta, and Jonathan Berant. 2023. Analyzing transformers in embedding space. In *Proceedings of ACL 2023*, pages 16124–16170.

Olaf Dimigen, Werner Sommer, Annette Hohlfeld, Arthur M Jacobs, and Reinhold Kliegl. 2011. Coregistration of eye movements and eeg in natural reading: analyses and review. *Journal of experimental psychology: General*, 140(4):552.

Chris Dyer, Adhiguna Kuncoro, Miguel Ballesteros, and Noah A. Smith. 2016. Recurrent neural network grammars. In *Proceedings of NAACL*. Association for Computational Linguistics.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of EMNLP-IJCNLP 2019*, pages 55–65.

Kara D Federmeier, Edward W Wlotko, Esmeralda De Ochoa-Dewald, and Marta Kutas. 2007. Multiple effects of sentential constraint on word processing. *Brain Research*, 1146:75–84.

Kenneth I Forster, Christine Guerrera, and Lisa Elliot. 2009. The maze task: measuring forced incremental sentence processing time. *Behavior research methods*, 41(1):163–171.

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

Stefan L Frank, Irene Fernandez Monsalve, Robin L Thompson, and Gabriella Vigliocco. 2013. Reading time data for evaluating broad-coverage models of english sentence processing. *Behavior research methods*, 45:1182–1190.

Stefan L. Frank, Leun J. Otten, Giulia Galli, and Gabriella Vigliocco. 2015. The erp response to the amount of information conveyed by words in sentences. *Brain and Language*, 140:1–11.

Richard Futrell, Edward Gibson, and Roger P. Levy. 2020. Lossy-Context Surprisal: An Information-Theoretic Model of Memory Effects in Sentence Processing. *Journal of Cognitive Science*.

Richard Futrell, Edward Gibson, Harry J Tily, Idan Blank, Anastasia Vishnevetsky, Steven T Piantadosi, and Evelina Fedorenko. 2021. The natural stories corpus: a reading-time corpus of english texts containing rare syntactic constructions. *Language Resource and Evaluation*, 55(1):63–77.

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 EMNLP 2024*, pages 18556–18572.

Mario Giulianelli, Andreas Opedal, and Ryan Cotterell. 2024b. Generalized measures of anticipation and responsivity in online language processing. In *Findings of EMNLP 2024*, pages 11648–11669.

Aaron Gokaslan and Vanya Cohen. 2019. Openwebtext corpus. `http://Skylion007.github.io/OpenWebTextCorpus`.

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.

Alex Graves. 2016. Adaptive computation time for recurrent neural networks. *arXiv preprint*.

Michael Hahn, Richard Futrell, Roger Levy, and Edward Gibson. 2022. A resource-rational model of human processing of recursive linguistic structure. *Proceedings of the National Academy of Sciences of the United States of America*, 119(43):e2122602119.

John Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of NAACL 2001*, pages 159–166.

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

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the sixth workshop on statistical machine translation*, pages 187–197.

Nora Hollenstein, Jonathan Rotsztejn, Marius Troendle, Andreas Pedroni, Ce Zhang, and Nicolas Langer. 2018. ZuCo, a simultaneous EEG and eye-tracking resource for natural sentence reading. *Scientific Data*, 5:180291.

Eghbal A Hosseini, Martin Schrimpf, Yian Zhang, Samuel Bowman, Noga Zaslavsky, and Evelina Fedorenko. 2024. Artificial neural network language models predict human brain responses to language even after a developmentally realistic amount of training. *Neurobiology of Language*, 5(1):43–63.

Jennifer Hu, Michael A Lepori, and Michael Franke. 2025. Signatures of human-like processing in transformer forward passes. *arXiv preprint arXiv:2504.14107*.

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.

Ryan J Hubbard, Joost Rommers, Cassandra L Jacobs, and Kara D Federmeier. 2019. Downstream behavioral and electrophysiological consequences of word prediction on recognition memory. *Frontiers in human neuroscience*, 13:291.

Marcel A. Just and Patricia A. Carpenter. 1980. A theory of reading: From eye fixations to comprehension. *Journal of Psychological Review*.

Yiğitcan Kaya, Sanghyun Hong, and Tudor Dumitras. 2019. Shallow-Deep Networks: Understanding and mitigating network overthinking. In *Proceedings of ICML 2019*.

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

Nikita Kitaev, Steven Cao, and Dan Klein. 2019. Multilingual constituency parsing with self-attention and pre-training. In *Proceedings of ACL 2019*, pages 3499–3505.

Nikita Kitaev and Dan Klein. 2018. Constituency parsing with a self-attentive encoder. In *Proceedings of ACL 2018*, pages 2676–2686.

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.

Marta Kutas and Kara D Federmeier. 2011. Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP). *Annual review of psychology*, 62(1):621–647.

Ellen F Lau, Colin Phillips, and David Poeppel. 2008. A cortical network for semantics: (de)constructing the N400. *Nature Reviews Neuroscience*, 9(12):920–933.

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

Richard L Lewis and Shravan Vasishth. 2005. An activation-based model of sentence processing as skilled memory retrieval. *Cogn. Sci.*, 29(3):375–419.

Falk Lieder and Thomas L Griffiths. 2019. Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, 43(e1):e1.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of EMNLP 2022*, pages 9019–9052.

David Marr. 1982. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. Henry Holt and Co., Inc., USA.

Kate McCurdy and Michael Hahn. 2024. Lossy context surprisal predicts task-dependent patterns in relative clause processing. In *Proceedings of CoNLL 2024*, pages 36–45.

Clara Meister, Mario Giulianelli, and Tiago Pimentel. 2024. Towards a similarity-adjusted surprisal theory. In *Proceedings of EMNLP 2024*, pages 16485–16498.

Clara Meister, Tiago Pimentel, Thomas Clark, Ryan Cotterell, and Roger Levy. 2022. Analyzing wrap-up effects through an information-theoretic lens. In *Proceedings of ACL 2022*, pages 20–28.

Danny Merkx and Stefan L. Frank. 2021. Human sentence processing: Recurrence or attention? In *Proceedings of CMCL*, pages 12–22.

James Michaelov, Catherine Arnett, and Ben Bergen. 2024a. Revenge of the fallen? recurrent models match transformers at predicting human language comprehension metrics. In *COLM 2024*.

James A. Michaelov, Megan D. Bardolph, Cyma K. Van Petten, Benjamin K. Bergen, and Seana Coulson. 2024b. Strong Prediction: Language Model Surprisal Explains Multiple N400 Effects. *Neurobiology of Language*, 5(1):107–135.

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

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

Samer Nour Eddine, Trevor Brothers, Lin Wang, Michael Spratling, and Gina R Kuperberg. 2024. A predictive coding model of the N400. *Cognition*, 246(105755):105755.

Byung-Doh Oh and William Schuler. 2023a. Transformer-based language model surprisal predicts human reading times best with about two billion training tokens. In *Findings of EMNLP 2023*, pages 1915–1921.

Byung-Doh Oh and William Schuler. 2023b. 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, Shisen Yue, and William Schuler. 2024. Frequency explains the inverse correlation of large language models' size, training data amount, and surprisal's fit to reading times. In *Proceedings of EACL 2024*, pages 2644–2663.

Andreas Opedal, Eleanor Chodroff, Ryan Cotterell, and Ethan Wilcox. 2024. On the role of context in reading time prediction. In *Proceedings of EMNLP 2024*, pages 3042–3058.

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

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

Keith Rayner. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological bulletin*, 124(3):372–422.

Keith Rayner and Charles Clifton, Jr. 2009. Language processing in reading and speech perception is fast and incremental: implications for event-related potential research. *Biological Psychology*, 80(1):4–9.

Keith Rayner, Gretchen Kambe, and Susan A. Duffy. 2000. The Effect of Clause Wrap-Up on Eye Movements during Reading. *Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology*, 53(4):1061–1080.

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.

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):e2105646118.

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

Cory Shain, Idan A Blank, Evelina Fedorenko, Edward Gibson, and William Schuler. 2022. Robust effects of working memory demand during naturalistic language comprehension in language-selective cortex. *Journal of Neuroscience*, 42(39):7412–7430.

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.

Noam Siegelman, Sascha Schroeder, Cengiz Acartürk, Hee-Don Ahn, Svetlana Alexeeva, Simona Amenta, Raymond Bertram, Rolando Bonandrini, Marc Brysbaert, Daria Chernova, et al. 2022. Expanding horizons of cross-linguistic research on reading: The multilingual eye-movement corpus (meco). *Behavior research methods*, 54(6):2843–2863.

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.

Jakub M Szewczyk and Kara D Federmeier. 2022. Context-based facilitation of semantic access follows both logarithmic and linear functions of stimulus probability. *Journal of Memory and Language*, 123(104311):104311.

Jakub M Szewczyk, Emily N Mech, and Kara D Federmeier. 2022. The power of "good": Can adjectives rapidly decrease as well as increase the availability of the upcoming noun? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 48(6):856–875.

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

Mariya Toneva and Leila Wehbe. 2019. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). *NeurIPS 2019*, pages 14928–14938.

Pranali Vani, Ethan Gotlieb Wilcox, and Roger Levy. 2021. Using the interpolated maze task to assess incremental processing in english relative clauses. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 43(43).

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.

Andrea Gregor de Varda, Marco Marelli, and Simona Amenta. 2024. Cloze probability, predictability ratings, and computational estimates for 205 english sentences, aligned with existing EEG and reading time data. *Behavior Research Methods*, 56(5):5190–5213.

Shravan Vasishth, Katja Suckow, Richard L Lewis, and Sabine Kern. 2010. Short-term forgetting in sentence comprehension: Crosslinguistic evidence from verb-final structures. *Language and Cognitive Processes*, 25(4):533–567.

Andreas Waldis, Yotam Perlitz, Leshem Choshen, Yufang Hou, and Iryna Gurevych. 2024. Holmes a benchmark to assess the linguistic competence of language models. *TACL*, 12:1616–1647.

Daphne Wang, Mehrnoosh Sadrzadeh, Miloš Stanojević, Wing-Yee Chow, and Richard Breheny. 2024. How can large language models become more human? In *Proceedings of CMCL 2024*, pages 166–176.

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 ACL 2024*, pages 15366–15394.

Ethan Wilcox, Clara Meister, Ryan Cotterell, and Tiago Pimentel. 2023a. Language model quality correlates with psychometric predictive power in multiple languages. In *Proceedings of EMNLP 2023*, pages 7503–7511.

Ethan Wilcox, Clara Meister, Ryan Cotterell, and Tiago Pimentel. 2023b. Language model quality correlates with psychometric predictive power in multiple languages. In *Proceedings of EMNLP 2023*, pages 7503–7511.

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

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*, pages 1707–1713.

Ethan Gotlieb Wilcox, Michael Hu, Aaron Mueller, Tal Linzen, Alex Warstadt, Leshem Choshen, Chengxu Zhuang, Ryan Cotterell, and Adina Williams. 2024. Bigger is not always better: The importance of human-scale language modeling for psycholinguistics. *PsyArXiv*.

Naoko Witzel, Jeffrey Witzel, and Kenneth Forster. 2012. Comparisons of online reading paradigms: eye tracking, moving-window, and maze. *Journal of Psycholinguistic Research*, 41(2):105–128.

Edward W Wlotko and Kara D Federmeier. 2012. So that's what you meant! event-related potentials reveal multiple aspects of context use during construction of message-level meaning. *Neuroimage*, 62(1):356–366.

Ryo Yoshida, Hiroshi Noji, and Yohei Oseki. 2021. Modeling human sentence processing with left-corner recurrent neural network grammars. In *Proceedings of EMNLP 2021*, pages 2964–2973.

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.

Wangchunshu Zhou, Canwen Xu, Tao Ge, Julian McAuley, Ke Xu, and Furu Wei. 2020. BERT loses patience: Fast and robust inference with early exit. *NeurIPS 2020*, abs/2006.04152.

A Details on psychometric predictive power

Here, we explain the ΔLL (PPP) score more formally. Recall that we quantify the predictive contribution of word-by-word surprisal to human cognitive responses using a log-likelihood-based score, ΔLL. Let $\boldsymbol{w} = [w_1, \ldots, w_n]^\top$ denote a sequence of words, and let $\boldsymbol{y} = [y_1, \ldots, y_n]^\top$ be the corresponding word-by-word human cognitive costs, e.g., reading times.

For each word w_i, we compute a surprisal value $s(w_i) \in \mathbb{R}$, and a vector of baseline linguistic features $b(w_i) \in \mathbb{R}^d$. We consider two linear models: a *full model* using both surprisal and baseline features, and a *reduced model* using only baseline features.

In both cases, we estimate the coefficients using ordinary least squares (OLS). The full model solves:

$$\hat{\phi}_s, \hat{\boldsymbol{\phi}}_b = \arg\min_{\phi_s, \boldsymbol{\phi}_b} \sum_{i=1}^{n} \left(y_i - \phi_s s(w_i) - \boldsymbol{\phi}_b^\top b(w_i) \right)^2 .$$

We assume homoscedastic Gaussian noise and evaluate the total log-likelihood of the fitted model using the maximum likelihood estimate of the variance:

$$\mathrm{LL}_{\text{full}} = -\frac{n}{2} \log(2\pi \hat{\sigma}^2_{\text{full}}) - \frac{n}{2}, \quad \text{where} \quad \hat{\sigma}^2_{\text{full}} = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \hat{\phi}_s s(w_i) - \hat{\boldsymbol{\phi}}_b^\top b(w_i) \right)^2 .$$

The reduced model is obtained by regressing y on baseline features alone:

$$\hat{\boldsymbol{\phi}}'_b = \arg\min_{\boldsymbol{\phi}_b} \sum_{i=1}^{n} \left(y_i - \boldsymbol{\phi}_b^\top b(w_i) \right)^2 ,$$

with corresponding log-likelihood:

$$\mathrm{LL}_{\text{baseline}} = -\frac{n}{2} \log(2\pi \hat{\sigma}^2_{\text{baseline}}) - \frac{n}{2}, \quad \text{where} \quad \hat{\sigma}^2_{\text{baseline}} = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \hat{\boldsymbol{\phi}}_b'^\top b(w_i) \right)^2 .$$

We define the *psychometric predictive power* of surprisal as the difference in log-likelihood between the full and reduced models:

$$\Delta\mathrm{LL} = \mathrm{LL}_{\text{full}} - \mathrm{LL}_{\text{baseline}} = \frac{n}{2} \log \left(\frac{\hat{\sigma}^2_{\text{baseline}}}{\hat{\sigma}^2_{\text{full}}} \right) .$$

This ΔLL value quantifies the contribution of surprisal to predicting human sentence processing data, with higher values indicating stronger predictive power.

B Models

Table 4 shows the exact models we used.

C Details on cross-lingual experiments

We target FPGD data from MECO in 13 languages (Siegelman et al., 2022) using five multilingual XGLMs (Lin et al., 2022) (564M, 1.7B, 2.9B, 4.5B, 7.5B). For the German part, we used monolingual German models in Table 4. We analyze the ΔLL of surprisal from their internal layers. Due to the unavailability of tuned-lenses for XGLMs, only logit-lenses were used. Figure 5 shows the relationship between paramter numbers and ΔLL in each language.

Model	URL	#params
GPT2	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-14m-deduped	https://huggingface.co/EleutherAI/pythia-14m-deduped	14M
Pythia-31m-deduped	https://huggingface.co/EleutherAI/pythia-31m-deduped	31M
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
Llama-3.1-8B	https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct	8B
Llama-3.1-70B	https://huggingface.co/meta-llama/Llama-3.1-70B-Instruct	70B
Qwen2.5-0.5B	https://huggingface.co/Qwen/Qwen2.5-0.5B	500M
Qwen2.5-1.5B	https://huggingface.co/Qwen/Qwen2.5-1.5B	1.5B
Qwen2.5-3B	https://huggingface.co/Qwen/Qwen2.5-3B	3B
Qwen2.5-7B	https://huggingface.co/Qwen/Qwen2.5-7B	7B
Qwen2.5-14B	https://huggingface.co/Qwen/Qwen2.5-14B	14B
Qwen2.5-32B	https://huggingface.co/Qwen/Qwen2.5-32B	32B
Qwen2.5-72B	https://huggingface.co/Qwen/Qwen2.5-72B	72B
XGLM-564M	https://huggingface.co/facebook/xglm-564M	564M
XGLM-1.7B	https://huggingface.co/facebook/xglm-1.7B	1.7B
XGLM-2.9B	https://huggingface.co/facebook/xglm-2.9B	2.9B
XGLM-4.5B	https://huggingface.co/facebook/xglm-4.5B	4.5B
XGLM-7.5B	https://huggingface.co/facebook/xglm-7.5B	7.5B
German LMs	https://huggingface.co/benjamin/gerpt2	124M
	https://huggingface.co/benjamin/gerpt2-large	774M
	https://huggingface.co/malteos/gpt2-xl-wechsel-german	1.5B

Table 4: LM details

	coef	std err	t	P> |t|	[0.025	0.975]
Intercept	0.0125	0.001	20.725	0.000	0.011	0.014
stimuli[T.Federmeier+, 2007]	0.0040	0.001	7.520	0.000	0.003	0.005
stimuli[T.Fillers]	0.0002	0.000	0.570	0.569	-0.001	0.001
stimuli[T.Hubbard+, 2019]	-0.0079	0.001	-14.755	0.000	-0.009	-0.007
stimuli[T.Michaelov+, 2024]	-0.0068	0.001	-12.666	0.000	-0.008	-0.006
stimuli[T.NS]	-0.0069	0.000	-14.314	0.000	-0.008	-0.006
stimuli[T.S&F,2022]	-0.0075	0.001	-13.955	0.000	-0.009	-0.006
stimuli[T.Szewczyk+, 2022]	-0.0033	0.001	-6.076	0.000	-0.004	-0.002
stimuli[T.UCL]	0.0090	0.000	22.701	0.000	0.008	0.010
stimuli[T.W&F,2012]	-0.0079	0.001	-14.816	0.000	-0.009	-0.007
stimuli[T.ZuCO]	0.0036	0.000	8.735	0.000	0.003	0.004
model[T.Llama-3.1-8B]	0.0010	0.001	1.707	0.088	-0.000	0.002
model[T.Qwen2.5-0.5B]	-0.0004	0.001	-0.689	0.491	-0.002	0.001
model[T.Qwen2.5-1.5B]	-0.0007	0.001	-1.101	0.271	-0.002	0.001
model[T.Qwen2.5-14B]	-0.0004	0.001	-0.822	0.411	-0.001	0.001
model[T.Qwen2.5-32B]	0.0002	0.000	0.473	0.637	-0.001	0.001
model[T.Qwen2.5-3B]	-0.0012	0.001	-2.129	0.033	-0.002	-9.41e-05
model[T.Qwen2.5-72B]	0.0003	0.000	0.731	0.465	-0.001	0.001
model[T.Qwen2.5-7B]	-0.0003	0.001	-0.481	0.630	-0.001	0.001
model[T.gpt2]	0.0010	0.001	1.568	0.117	-0.000	0.002
model[T.gpt2-large]	0.0020	0.000	4.266	0.000	0.001	0.003
model[T.gpt2-medium]	0.0008	0.001	1.242	0.214	-0.000	0.002
model[T.gpt2-xl]	0.0023	0.000	5.367	0.000	0.001	0.003
model[T.opt-1.3b]	0.0026	0.001	5.084	0.000	0.002	0.004
model[T.opt-125m]	0.0024	0.001	3.778	0.000	0.001	0.004
model[T.opt-13b]	0.0028	0.001	5.301	0.000	0.002	0.004
model[T.opt-2.7b]	0.0028	0.001	4.902	0.000	0.002	0.004
model[T.opt-30b]	0.0023	0.001	4.513	0.000	0.001	0.003
model[T.opt-6.7b]	0.0027	0.000	5.639	0.000	0.002	0.004
model[T.opt-66b]	0.0027	0.000	5.721	0.000	0.002	0.004
model[T.pythia-1.4b-deduped]	0.0010	0.001	1.992	0.046	1.65e-05	0.002
model[T.pythia-12b-deduped]	0.0023	0.000	4.920	0.000	0.001	0.003
model[T.pythia-14m]	-0.0017	0.001	-1.508	0.132	-0.004	0.000
model[T.pythia-160m-deduped]	-0.0015	0.001	-2.352	0.019	-0.003	-0.000
model[T.pythia-1b-deduped]	0.0010	0.001	1.367	0.172	-0.000	0.002
model[T.pythia-1b-deduped-v0]	0.0040	0.002	2.415	0.016	0.001	0.007
model[T.pythia-2.8b-deduped]	0.0016	0.000	3.349	0.001	0.001	0.003
model[T.pythia-31m]	-0.0015	0.001	-1.391	0.164	-0.004	0.001
model[T.pythia-410m-deduped]	0.0008	0.001	1.476	0.140	-0.000	0.002
model[T.pythia-6.9b-deduped]	0.0018	0.000	3.826	0.000	0.001	0.003
model[T.pythia-70m-deduped]	-0.0016	0.001	-1.911	0.056	-0.003	4.05e-05
measure[T.FPGD]	0.0053	0.000	10.675	0.000	0.004	0.006
measure[T.N400]	-0.0069	0.001	-13.720	0.000	-0.008	-0.006
measure[T.MAZE]	-0.0160	0.001	-28.315	0.000	-0.017	-0.015
method[T.tuned-lens]	0.0003	0.000	1.365	0.172	-0.000	0.001
normalized_layer	-0.0076	0.001	-12.601	0.000	-0.009	-0.006
measure[T.FPGD]:layer	-0.0034	0.001	-4.242	0.000	-0.005	-0.002
measure[T.N400]:layer	0.0100	0.001	14.084	0.000	0.009	0.011
measure[T.MAZE]:layer	0.0593	0.001	61.971	0.000	0.057	0.061

(a) All data

	coef	std err	t	P> |t|	[0.025	0.975]
Intercept	-0.0002	0.000	-0.971	0.331	-0.001	0.000
stimuli[T.Federmeier+, 2007]	0.0394	0.000	168.408	0.000	0.039	0.040
stimuli[T.Fillers]	0.0011	0.000	7.379	0.000	0.001	0.001
stimuli[T.Hubbard+, 2019]	0.0275	0.000	117.432	0.000	0.027	0.028
stimuli[T.Michaelov+, 2024]	0.0286	0.000	122.213	0.000	0.028	0.029
stimuli[T.NS]	-0.0011	0.000	-6.066	0.000	-0.001	-0.001
stimuli[T.S&F,2022]	0.0279	0.000	119.263	0.000	0.027	0.028
stimuli[T.Szewczyk+, 2022]	0.0321	0.000	137.295	0.000	0.032	0.033
stimuli[T.UCL]	0.0029	0.000	19.160	0.000	0.003	0.003
stimuli[T.W&F,2012]	0.0275	0.000	117.292	0.000	0.027	0.028
stimuli[T.ZuCO]	0.0274	0.000	165.961	0.000	0.027	0.028
model[T.Llama-3.1-8B]	6.679e-05	0.000	0.299	0.765	-0.000	0.001
model[T.Qwen2.5-0.5B]	-0.0002	0.000	-0.667	0.505	-0.001	0.000
model[T.Qwen2.5-1.5B]	9.912e-05	0.000	0.424	0.672	-0.000	0.001
model[T.Qwen2.5-14B]	3.475e-05	0.000	0.178	0.859	-0.000	0.000
model[T.Qwen2.5-32B]	-2.221e-05	0.000	-0.123	0.902	-0.000	0.000
model[T.Qwen2.5-3B]	-0.0003	0.000	-1.455	0.146	-0.001	0.000
model[T.Qwen2.5-72B]	-0.0004	0.000	-2.508	0.012	-0.001	-9.3e-05
model[T.Qwen2.5-7B]	-0.0005	0.000	-2.061	0.039	-0.001	-2.36e-05
model[T.gpt2]	-0.0003	0.000	-1.304	0.192	-0.001	0.000
model[T.gpt2-large]	3.558e-05	0.000	0.199	0.842	-0.000	0.000
model[T.gpt2-medium]	-0.0005	0.000	-2.208	0.027	-0.001	-6.13e-05
model[T.gpt2-xl]	0.0005	0.000	2.718	0.007	0.000	0.001
model[T.opt-1.3b]	0.0004	0.000	1.794	0.073	-3.3e-05	0.001
model[T.opt-125m]	-0.0001	0.000	-0.411	0.681	-0.001	0.000
model[T.opt-13b]	0.0004	0.000	1.747	0.081	-4.41e-05	0.001
model[T.opt-2.7b]	0.0005	0.000	2.341	0.019	8.51e-05	0.001
model[T.opt-30b]	0.0002	0.000	1.040	0.298	-0.000	0.001
model[T.opt-6.7b]	0.0001	0.000	0.669	0.503	-0.000	0.000
model[T.opt-66b]	0.0005	0.000	2.709	0.007	0.000	0.001
model[T.pythia-1.4b-deduped]	-5.816e-05	0.000	-0.293	0.770	-0.000	0.000
model[T.pythia-12b-deduped]	0.0005	0.000	2.831	0.005	0.000	0.001
model[T.pythia-14m]	-0.0018	0.000	-4.125	0.000	-0.003	-0.001
model[T.pythia-160m-deduped]	-0.0008	0.000	-3.296	0.001	-0.001	-0.000
model[T.pythia-1b-deduped]	-0.0004	0.000	-1.331	0.183	-0.001	0.000
model[T.pythia-1b-deduped-v0]	0.0002	0.001	0.294	0.769	-0.001	0.001
model[T.pythia-2.8b-deduped]	-0.0001	0.000	-0.646	0.519	-0.000	0.000
model[T.pythia-31m]	-0.0018	0.000	-4.200	0.000	-0.003	-0.001
model[T.pythia-410m-deduped]	-0.0005	0.000	-2.451	0.014	-0.001	-9.76e-05
model[T.pythia-6.9b-deduped]	0.0003	0.000	1.807	0.071	-2.81e-05	0.001
model[T.pythia-70m-deduped]	-0.0013	0.000	-4.092	0.000	-0.002	-0.001
measure[T.FPGD]	0.0053	0.000	27.964	0.000	0.005	0.006
measure[T.N400]	-0.0301	0.000	-124.125	0.000	-0.031	-0.030
measure[T.MAZE]	0.0007	0.000	3.082	0.002	0.000	0.001
method[T.tuned-lens]	-0.0003	8.25e-05	-3.448	0.001	-0.000	-0.000
normalized_layer	0.0015	0.000	6.663	0.000	0.001	0.002
measure[T.FPGD]:layer	-0.0088	0.000	-29.286	0.000	-0.009	-0.008
measure[T.N400]:layer	0.0046	0.000	16.990	0.000	0.004	0.005
measure[T.MAZE]:layer	0.0030	0.000	8.430	0.000	0.002	0.004

(b) Only with sentence/clause-final tokens

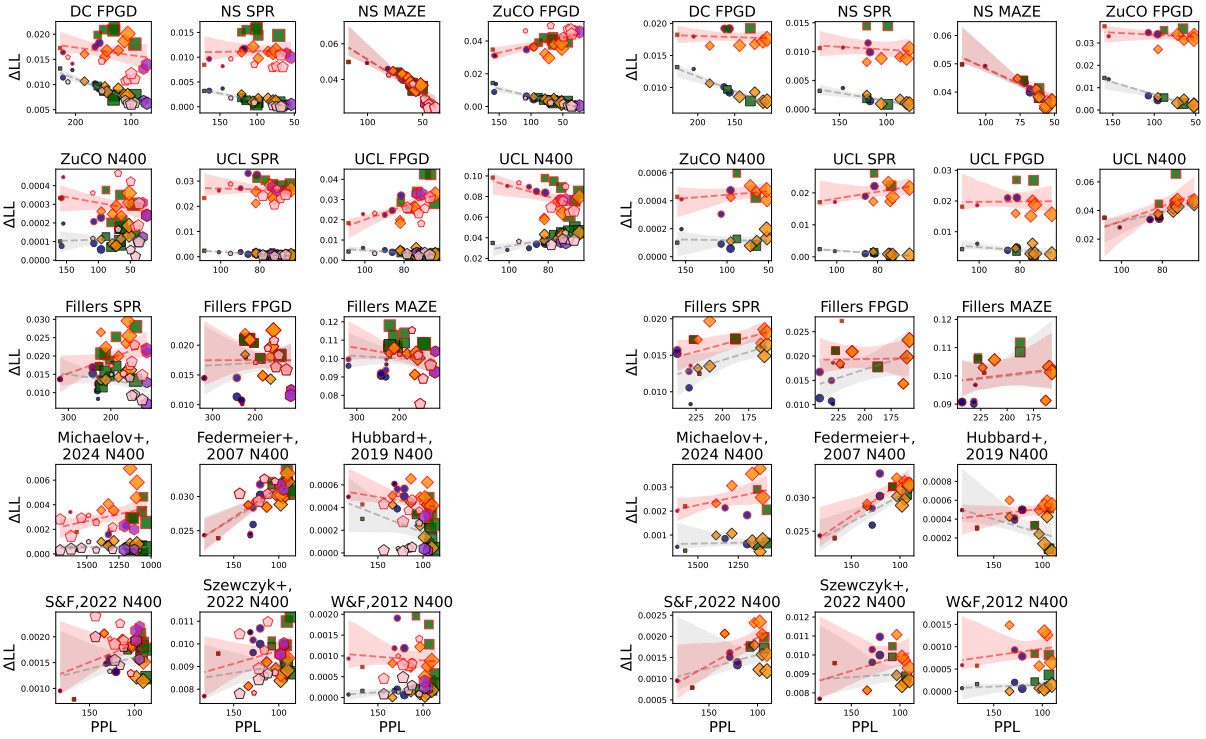
Table 5: Obtained coefficients in Section 6.2

Features	coef	std err	t	P> |t|	[0.025	0.975]
Intercept	0.0801	0.430	0.186	0.852	-0.764	0.924
model[T.gpt2-large]	0.1510	0.061	2.490	0.013	0.032	0.270
model[T.gpt2-xl]	0.1647	0.061	2.716	0.007	0.046	0.284
model[T.opt-1.3b]	0.1293	0.061	2.132	0.033	0.010	0.248
model[T.opt-125m]	0.0182	0.061	0.301	0.764	-0.101	0.137
model[T.opt-6.7b]	0.1898	0.061	3.131	0.002	0.071	0.309
model[T.pythia-1.4b-deduped]	0.0754	0.061	1.243	0.214	-0.043	0.194
model[T.pythia-12b-deduped]	0.1649	0.061	2.719	0.007	0.046	0.284
model[T.pythia-160m-deduped]	0.1237	0.061	2.040	0.041	0.005	0.243
model[T.pythia-1b-deduped-v0]	0.0695	0.061	1.146	0.252	-0.049	0.188
model[T.pythia-2.8b-deduped]	0.1212	0.061	1.998	0.046	0.002	0.240
model[T.pythia-410m-deduped]	0.0607	0.061	1.001	0.317	-0.058	0.180
model[T.pythia-6.9b-deduped]	0.1534	0.061	2.529	0.011	0.035	0.272
model[T.pythia-70m-deduped]	-0.0478	0.061	-0.788	0.431	-0.167	0.071
pos[T.$]	-0.1911	0.776	-0.246	0.805	-1.711	1.329
pos[T.,]	-4.1266	0.985	-4.188	0.000	-6.058	-2.195
pos[T..]	-2.5070	1.327	-1.890	0.059	-5.107	0.093
pos[T.CC]	-0.7410	0.431	-1.719	0.086	-1.586	0.104
pos[T.CD]	-1.6381	0.443	-3.699	0.000	-2.506	-0.770
pos[T.DT]	-0.9570	0.428	-2.238	0.025	-1.795	-0.119
pos[T.EX]	-1.5530	0.489	-3.176	0.001	-2.511	-0.595
pos[T.FW]	4.2889	0.580	7.395	0.000	3.152	5.426
pos[T.IN]	-1.1206	0.427	-2.625	0.009	-1.957	-0.284
pos[T.JJ]	-1.4790	0.427	-3.461	0.001	-2.317	-0.641
pos[T.JJR]	-1.6572	0.463	-3.580	0.000	-2.565	-0.750
pos[T.JJS]	-2.1918	0.488	-4.489	0.000	-3.149	-1.235
pos[T.MD]	-1.6324	0.435	-3.749	0.000	-2.486	-0.779
pos[T.NN]	-2.0801	0.427	-4.876	0.000	-2.916	-1.244
pos[T.NNP]	-1.0809	0.428	-2.523	0.012	-1.921	-0.241
pos[T.NNPS]	-5.2269	0.484	-10.804	0.000	-6.175	-4.279
pos[T.NNS]	-2.1715	0.428	-5.071	0.000	-3.011	-1.332
pos[T.PDT]	-2.2381	0.535	-4.183	0.000	-3.287	-1.189
pos[T.POS]	-2.0896	0.449	-4.651	0.000	-2.970	-1.209
pos[T.PRP]	-1.0883	0.430	-2.529	0.011	-1.932	-0.245
pos[T.PRP$]	-1.4344	0.435	-3.296	0.001	-2.287	-0.581
pos[T.RB]	-1.7242	0.428	-4.026	0.000	-2.564	-0.885
pos[T.RBR]	-1.8143	0.482	-3.761	0.000	-2.760	-0.869
pos[T.RBS]	-3.0363	0.529	-5.743	0.000	-4.073	-2.000
pos[T.RP]	-1.4753	0.456	-3.235	0.001	-2.369	-0.582
pos[T.SYM]	2.6044	0.901	2.889	0.004	0.838	4.371
pos[T.TO]	-0.6908	0.431	-1.602	0.109	-1.536	0.155
pos[T.UH]	-7.8480	0.759	-10.338	0.000	-9.336	-6.360
pos[T.VB]	-1.2679	0.429	-2.955	0.003	-2.109	-0.427
pos[T.VBD]	-1.6735	0.431	-3.885	0.000	-2.518	-0.829
pos[T.VBG]	-2.0877	0.433	-4.818	0.000	-2.937	-1.238
pos[T.VBN]	-1.7750	0.431	-4.118	0.000	-2.620	-0.930
pos[T.VBP]	-1.1485	0.432	-2.659	0.008	-1.995	-0.302
pos[T.VBZ]	-1.2311	0.431	-2.860	0.004	-2.075	-0.387
pos[T.WDT]	-1.2174	0.449	-2.714	0.007	-2.097	-0.338
pos[T.WP]	-1.0289	0.455	-2.264	0.024	-1.920	-0.138
pos[T.WP$]	0.1774	0.869	0.204	0.838	-1.525	1.880
pos[T.WRB]	-0.8454	0.456	-1.853	0.064	-1.740	0.049
has_punct[T.True]	-1.6396	0.038	-43.221	0.000	-1.714	-1.565
has_num[T.True]	2.3217	0.171	13.602	0.000	1.987	2.656
freq	0.0142	0.007	1.930	0.054	-0.000	0.029
length	0.3629	0.007	51.621	0.000	0.349	0.377
token_position_in_sentence	-0.0042	0.001	-3.958	0.000	-0.006	-0.002

Table 6: Obtained coefficients in Section 6.4. Note that word frequency and length are confounded (Pearson's $r = -0.7$), and once the length factor is excluded, the coefficient for the word frequency feature becomes significantly negative and has a high t-score. That is, a large decrease in regression error is generally associated with infrequent, long words.

	Logit-lens (ΔLL)					Tuned-lens (ΔLL)				
Measure	0-0.2	0.2-0.4	0.4-0.6	0.6-0.8	0.8-1	0-0.2	0.2-0.4	0.4-0.6	0.6-0.8	0.8-1
SPR	**24.51**	23.35	18.84	7.80	1.81	**15.78**	8.92	4.87	2.53	1.27
FPGD	22.11	**23.39**	22.83	15.77	6.53	**16.28**	14.48	11.87	9.47	5.57
SPGD	**228.29**	189.49	144.70	72.48	21.47	**68.43**	20.05	5.28	1.71	3.87
TOTAL	**201.36**	171.44	134.24	68.64	19.82	**67.39**	23.51	7.82	1.61	1.56
ELAN (125–175ms)	**0.75**	0.48	0.31	0.36	0.34	0.22	0.38	**0.77**	0.67	0.50
LAN (300–400ms)	**67.55**	51.86	35.96	17.45	8.17	**19.94**	2.74	1.72	5.02	8.76
N400 (300–500ms)	**56.86**	38.04	22.77	16.07	22.58	11.31	6.12	16.19	29.49	**37.11**
EPNP (400–600ms)	**75.94**	61.74	46.02	22.85	7.84	**29.31**	6.67	1.29	1.56	4.45
P600 (500–700ms)	**69.02**	55.91	41.88	22.60	10.37	**17.70**	3.30	1.80	5.83	11.36
PNP (600–700ms)	**59.75**	50.83	39.67	21.40	7.26	**25.38**	8.09	2.30	0.73	1.61

Table 7: Results for other human measures recorded on the UCL corpus. ΔLLs are multiplied by 1000.



(a) Logit-lens

(b) Tuned-lens

Figure 8: Scaling effect between ΔLL and PPL (measured on respective datasets with final layer), instead of model parameter counts, as adopted in Figure 3.

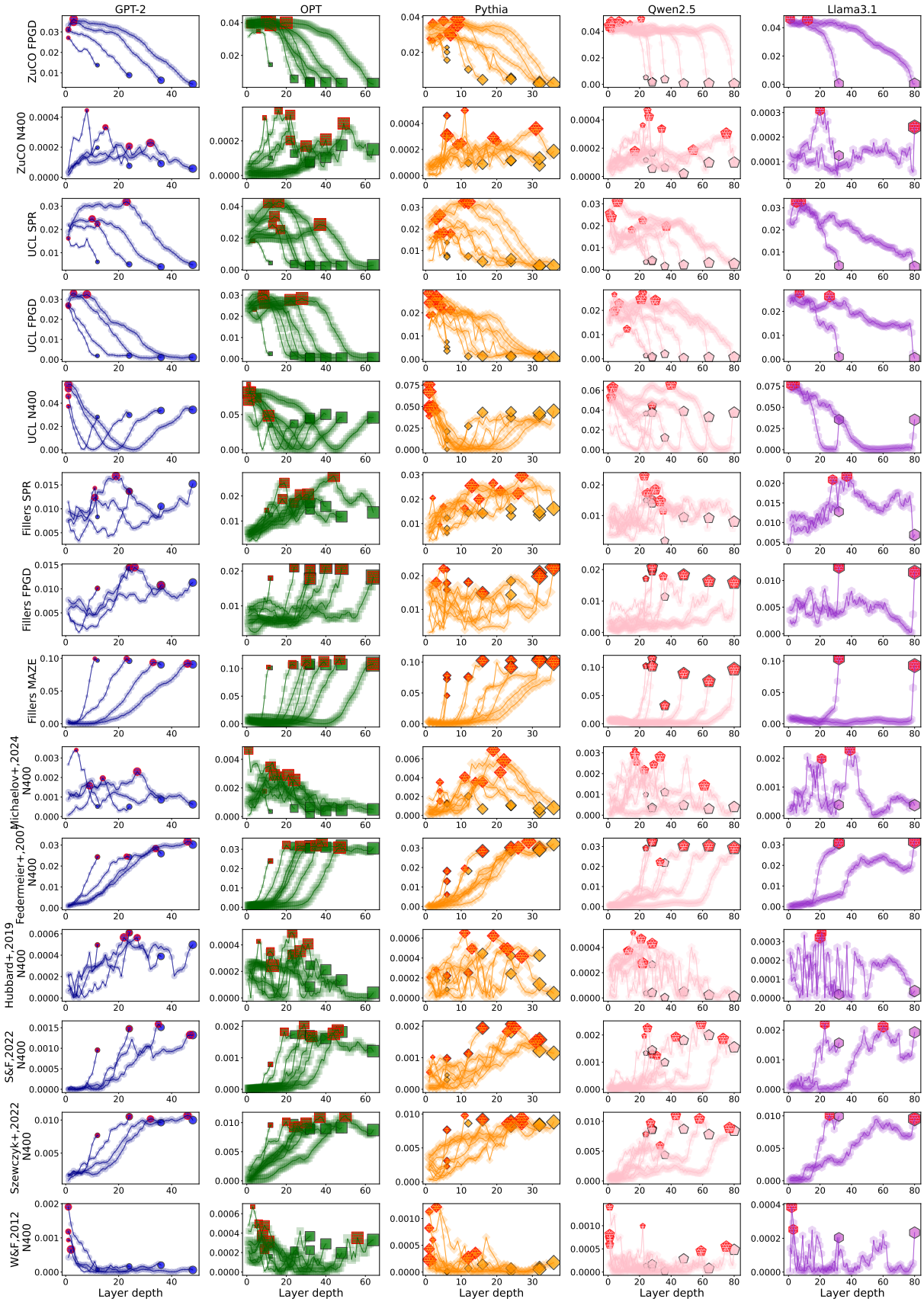


Figure 9: Visulaition of layer–ΔLL relationships (in addition to Figure 1)