Large Language Models are Human-Like Internally

Tatsuki Kuribayashi[1,2] Yohei Oseki[3] Souhaib Ben Taieb[1,4] Kentaro Inui[1,2,5] Timothy Baldwin[1,6]

[1] MBZUAI [2]Tohoku University [3]The University of Tokyo [4]University of Mons [5]RIKEN [6]The University of Melbourne

Prediction of upcoming words has been regarded as a key component of human online sentence processing. To give a proof-of-concept of this idea, human reading behavior has been compared with next-word probabilities obtained from language models (LMs). Specifically, these studies typically build on surprisal theory [1, 2] as a linking hypothesis — human cognitive loads, measured by, e.g., reading time, are compared with the logarithmic probabilities of a word in context computed by LMs, namely, surprisal $-\log(\text{word}|\text{context})$.

In this study, we extend the scope of such a surprisal-based cognitive modeling with LMs into their model internals, while existing studies have exclusively focused on the final probability output of LMs. Notably, neural LMs consist of a stack of layers, and the input representation gradually evolves throughout the layers. Some techniques to extract surprisal from internal layers have been developed in the mechanistic interpretability field in natural language processing (NLP). Given many layer options to compute surprisal, it is not obvious that the probability/representation from the final layer should be a counterpart for human measures. Particularly, if relatively "fast" human measures, such as first pass gaze duration, reflect the early stage of human sentence processing, their counterpart in LMs might be probability/representation in their early layers. Our extended scope into LMs' internals in cognitive modeling opens such new questions to bridge LMs' internal dynamics with human sentence processing. More specifically, in our experiments, we adopt a simple method called *logit-lens* [3] and its variant [4], which have been investigated in the LLM interpretability field. The logit-lens utilizes LM's prediction head $H : \mathbb{R}^d \rightarrow [0,1]^{|\mathcal{V}|}$ that maps d-dimensional representation in the final layer into a probability distribution over vocabulary $\mathcal{V}$. This head H is simply applied to internal layers of the model to obtain surprisal values (given residual connections that potentially avoid representational drift across layers, this empirically yields reasonable next-word probabilities).

We systematically evaluate the fit of the surprisal from internal layers with human measures, using 30 LMs and 15 datasets where human reading behavior/physiology data are recorded (Table 1). Following existing studies [5, 6], we evaluate the goodness-of-fit of the regression model to predict word-by-word human measures with baseline linguistic features and surprisal. We report the delta loglikelihood scores (ΔLL; psychometric predictive power) — the increase of loglikelihood before and after adding the surprisal feature to the regression model. We obtain several key findings:

1. Surprisal from internal layers typically better fits with human measures than the final layers previously focused on in existing studies. Cognitive alignment between the sentence processing of humans and LMs can be made in internal, earlier layers.
2. There are systematic tendencies between human measure types and their aligned LM layers; for example, first-pass gaze duration and self-paced reading time tend to align with early layers, while N400 and MAZE data align with middle or latter layers, on average. This may suggest the parallel between the real time-scale in human measures and that in the layer direction, if one admits that the former human measures aligned with earlier layers, such as gaze duration, are "fast" and the latter, such as N400, are "slow" ones.
3. Once relying on the best fitting layer, larger LMs tend to exhibit a comparable or better fit with human measures than smaller LMs. This overrides the previous observation relying only on the final layers — larger Transformer-based LMs show worse fit of their surprisal to human reading times [6] (Figure 1).

Table 1: 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. References for the datasets are omitted due to page limits.

		PPP logit-lens					PPP tuned-lens				
Stimuli	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
Dundee Corpus	FPGD	14.13	14.78	**14.92**	13.38	9.84	**17.10**	16.32	15.39	13.53	10.49
Natural Stories Corpus	SPR	**9.85**	9.75	8.44	5.68	2.67	**8.93**	7.03	5.11	3.44	2.33
	MAZE	1.18	3.00	5.69	12.06	**23.77**	9.70	17.56	24.15	32.86	**39.63**
ZuCO Corpus	FPGD	38.10	**38.13**	35.59	29.82	15.94	**30.48**	27.16	22.56	17.29	8.77
	N400	0.07	0.12	0.15	**0.18**	0.16	0.20	0.32	**0.34**	0.29	0.18
UCL Corpus	SPR	**22.88**	22.21	19.30	11.45	4.77	**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
	N400	**56.86**	38.04	22.77	16.07	22.58	11.31	6.12	16.19	29.49	**37.11**
Fillers in [7]	SPR	7.83	10.89	**14.39**	14.18	14.35	8.60	10.47	11.36	11.86	**13.33**
	FPGD	6.66	5.83	6.48	7.31	**10.36**	8.94	10.91	12.91	13.81	**14.00**
	MAZE	5.39	3.01	5.08	21.97	**60.89**	9.96	28.27	52.00	73.38	**88.64**
Michaelov+ 2024	N400	0.88	1.42	**1.91**	1.68	0.91	0.95	1.51	**1.70**	1.38	0.99
Federmeier+ 2007	N400	0.77	3.11	8.59	18.05	**25.80**	1.49	5.22	13.06	24.48	**28.71**
W&F 2012	N400	**0.35**	0.19	0.10	0.09	0.12	**0.51**	0.27	0.12	0.05	0.11
Hubbard+ 2019	N400	0.18	0.23	0.23	**0.25**	0.17	0.11	0.12	0.22	**0.36**	0.33
S&F 2022	N400	0.11	0.15	0.38	0.90	**1.40**	0.16	0.33	0.77	1.29	**1.42**
Szewczyk+ 2022	N400	1.21	2.91	4.43	6.40	**8.04**	2.12	3.58	5.52	8.10	**8.93**

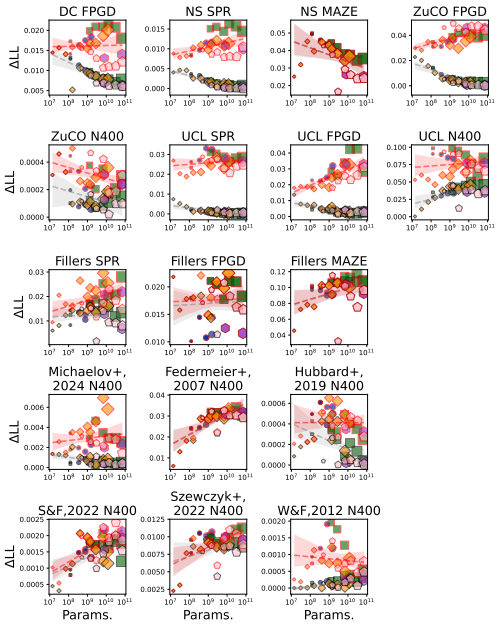

Figure 1: Relationship between LM parameter size (the larger the right side) and psychometric predictive power (the higher the better). The gray line relies on LM's last layer, and the red line relies on LM's best layer. The results with logit-lens.

References

[1] Levy, R. (2008). Expectation-based syntactic comprehension. *Journal of Cognition*, *106*(3), 1126–1177.

[2] Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*(3), 302–319.

[3] nostalgebraist. (2020). Interpreting GPT: The logit lens.

[4] Belrose, N., Furman, Z., Smith, L., Halawi, D., McKinney, L., Ostrovsky, I., Biderman, S., & Steinhardt, J. (2023). Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint.*

[5] Kuribayashi, T., Oseki, Y., Ito, T., Yoshida, R., Asahara, M., & Inui, K. (2021). Lower perplexity is not always human-like. *Proceedings of ACL-IJCNLP 2021*, 5203–5217.

[6] Oh, B.-D., & Schuler, W. (2023). Why does surprisal from larger transformer-based language models provide a poorer fit to human reading times? *TACL*, *11*, 336–350.

[7] Vasishth, S., Suckow, K., Lewis, R. L., & Kern, S. (2010). Short-term forgetting in sentence comprehension: Crosslinguistic evidence from verb-final structures. *Language and Cognitive Processes*, *25*(4), 533–567.