大規模言語モデルの浅い層が人間の速い言語処理を再現する

栗林樹生 [1] 大関洋平 [2] Souhaib Ben Taieb[1] 乾健太郎 [1,3,4] Timothy Baldwin[1,5]
[1]MBZUAI [2] 東京大学 [3] 東北大学 [4] 理化学研究所 [5] メルボルン大学
tatsuki.kuribayashi@mbzuai.ac.ae

概要

本研究では，大規模ニューラル言語モデルの中間層から得られる次単語予測確率が人間の読み活動・脳波データをうまく説明できることを報告する．これは，大きな言語モデルの計算する確率ほど人間の振る舞いから逸脱してしまうという既存の報告を覆すものであり，大規模言語モデルの認知的妥当性が過小評価されていたことを示唆する．さらに，人間の比較的速い反応（最初の視線停留など）は浅い層で，遅い反応（脳波 N400 など）は深い層で再現される傾向が見られ，言語モデル内部でのレイヤ方向の漸進的処理と，人間の異なるタイムスケールの反応との対応をとる新たな方向性を提示する．

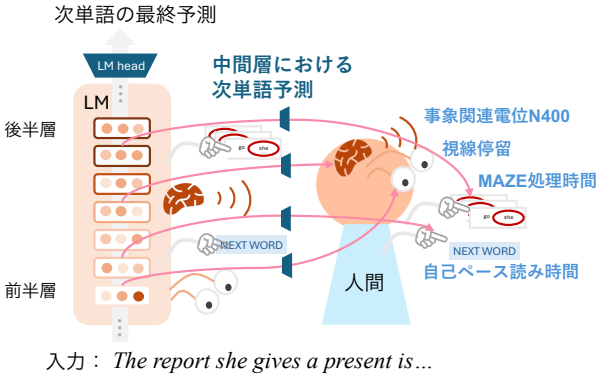

図 1 言語モデル内の異なる中間層が，人間の異なる読み活動・脳波データを説明する．中間層を含めた認知モデリングでは，大きな言語モデルも人間をうまく模倣する．

1 はじめに

1.1 背景

人間は，逐次的に文章を読んでいる間，様々な行動・生理的反応を見せる．例えば，しばらくある単語を見た後，前後の単語に視線を移したり，特徴的な脳活動を示したりする．このようなデータの説明は，人間の言語処理で何が計算されているかという問いに答える手がかりとなる．一つの仮説として，人間も（大雑把に言えば言語モデルと同じように）言語を処理している最中に，次の単語・情報を予測しており，予測しづらい箇所で認知的負荷が高くなる（視線が停留したり，特定の生理的反応が生じる），さらにその関係は文脈内での単語の対数確率 $-\log p(\text{単語} \mid \text{前方文脈})$ に線形であるという，サプライザル仮説があげられる [1, 2, 3, 4]．本研究はこの仮説を掘り下げるものである．

サプライザル仮説そのものは，計算のゴール・目的（将来の予測）についてのみ言及しており，それがどのような表現・アルゴリズムで計算されるかといったレベルには踏み込んでいない．当該領域では，このアルゴリズムの理解（例えば，本当に人間は木のような統語構造を扱って次単語予測をしているとみなせるのかなど）が次の関心となっている．幸い，自然言語処理分野は次単語確率・サプライザル を計算できる様々なアルゴリズム（逐次的パーザや言語モデルなど）を開発してきた．これらを仮説の実装として利用・改造し，どのモデルの計算するサプライザルが，人間に近い振る舞いを示すかが検証されている．そのモデル候補として，近年であれば，人間らしい流暢な言語を生成できる大規模言語モデルも含められてきた．

ここ数年の発見として，大きな言語モデルほど，得られる単語確率（サプライザル）は人間の読み活動から**乖離してしまう**という報告があげられてきた [5, 6, 7]．すなわち，大規模言語モデルが「驚く」箇所で，人間も驚いているわけではなく，むしろ GPT2-small 程度の小さなモデルの驚き方が最も人間らしいという帰結が得られている．多くの研究者は，これを踏まえ，なぜ大きな言語モデルほど人間から乖離していくのかという分析を始めているが [8, 9, 10, 11]，相補的な取り組みとして，我々は，**本当に大規模言語モデルから人間らしい予測・驚きが得られないのか？**という点について再訪する．

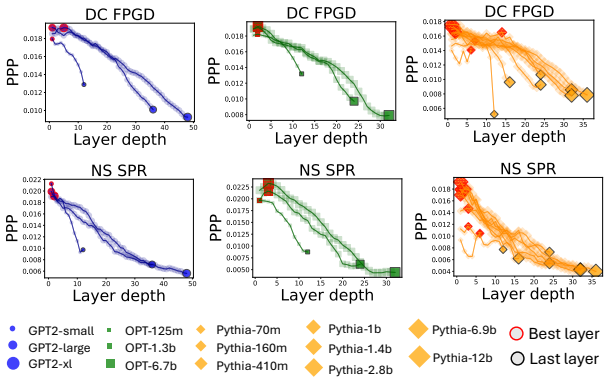

図 2　言語モデルの前半層が人間の読み活動データをうまく説明できる例．モデルシリーズごとの結果．横軸が言語モデルの層の位置，縦軸が人間データとの適合の良さ．

1.2　提案：言語モデルの中間予測

本研究ではとりわけ，ニューラル言語モデルを活用したサプライザル仮説の検証 [12, 13, 14, 15, 5] において暗黙に仮定される，**モデルの最終出力として得られる確率で認知モデリングをする**という設定について再検討する．そして実験では，人間を測定する指標に応じて，むしろ前半層や中間層から得られる，洗練される前の言語モデルの中間予測が人間のデータと適合することを示す．つまり，仮に人間を大規模言語モデルと見立てた場合，人間から即時的に観察される反応は，中間層の情報のようなものに対応すると考えれば[1)]，ある程度説明がつく（図 2）．

このような検証に至る動機として，そもそも計測されている人間の反応が，言語モデルの最終層のような，その語に対する処理が完了するまでの全ての処理・予測・判断を踏まえたものではない可能性をあげたい．人間に対する異なる測定指標はしばしば異なるタイムスケールに対応しており，例えば，N400 事象関連電位がおよそ 400ms 後に出現する前に，視線（first pass gaze duration; FPGD）はすでに次の単語に移っていることが多い [16, 17]．この観点では，FPGD は次の単語を迎え入れるまでの必要最小限の処理しか反映していないかもしれない．更に少々荒い議論でよければ，例えば，眼は情報が入るセンサーであり，そのような眼球・視線の（ある程度反射的な）振る舞いは言語モデルの入力層のようでもおかしくないのかもしれない．一般的には，速い指標は語彙などの浅い処理に，遅い指標は意味などの深い処理に紐づくという見方がされ [18, 19, 20]，このような同じ単語を処理している最中の異なる時間スケール・粒度の反応は，言語モデル側で考えれば，層方向の違いが候補にあがるだろう．さらに，人間の速い反応は先に計算される前半層で再現され，遅い反応はその後計算される後半層で再現されると説明できるのであれば．直観的である．本研究では，少なくとも FPGD, self-paced reading time (SPR), N400, MAZE などのよく用いられる人間の指標については，このような対応が概ね当てはまることを示す．また，大規模言語モデル内の多くの中間層では，これまで良いとされた小さい言語モデルと同等かそれ以上に人間の振る舞い・脳波を再現でき，これは大きい言語モデルほど人間から逸脱していくという既存の知見 [14, 5, 6, 7] を覆す．

1）　後半層の処理にあたるものは，認知モデリングの文脈で計測される，視線移動のような即時的な行動には反映されない，熟慮のようなものだと一旦思うことにする．

2　実験設定

2.1　認知モデリング

ある文章 $[w_0, w_1, \cdots, w_n]$ を逐次的に読んでいる最中に，人間が各単語 w_t に対して示す処理負荷を $c(w_t|\boldsymbol{w}_{<t})$ とする．この処理負荷は例えば，視線停留時間の長さであったり，特定の脳波のピークの大きさであったりする．同様の研究に従い，各単語のサプザイザル $-\log p(w_t|\boldsymbol{w}_{<t})$ でこの処理負荷を説明できると期待し，言語モデルを用いて近似する．ここでどのような言語モデル（本研究では層）を用いればよいかが関心である．単に $c(w_t|\boldsymbol{w}_{<t})$ と $-\log p(w_t|\boldsymbol{w}_{<t})$ の相関で評価してもよいが，より洗練された方法として，既存研究に基づき線形回帰モデルを用い，単語頻度や単語の長さといった素朴な要因を配慮したうえで，サプライザルが処理負荷の説明にどれだけ寄与するかを調べる [12, 14, 21, 15, 5]．具体的には，サプライザルを含めた回帰モデルと含めない回帰モデルで処理負荷を説明したときの対数尤度の差（psychometric predictive power; PPP）を報告し，PPP が大きいほど，用いたサプライザルが人間の処理負荷をうまく説明している．具体的な回帰式は付録 A に載せる．

2.2　言語モデルの中間予測

本研究では，言語モデルの中間層 $\boldsymbol{h}_l \in \mathbb{R}^d$ から，単語サプライザル $-\log p(w_t|w_{<t}; \boldsymbol{h}_l)$ を計算する．簡単な手法として，言語モデリングヘッドを中間層

表 1　全結果．各層の PPP は特定の層範囲（相対的）で，全モデルから得られた値を平均して報告している．例えば 0-20 は全モデルの最初の 20%のレイヤから得られた PPP を平均したものである．各値は 1000 倍されている．

		Logit-lens (PPP)					Tuned-lens (PPP)				
テキスト	指標	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 [22]	16.69	**17.90**	17.55	15.24	10.38	**17.10**	16.32	15.39	13.53	10.49
NS	SPR [23]	23.81	**23.92**	21.26	13.36	6.88	**16.88**	14.53	11.36	8.14	6.32
	MAZE [24]	1.66	4.32	8.36	18.50	**32.40**	9.70	17.56	24.15	32.86	**39.63**
UCL	SPR [25]	**24.81**	23.42	18.88	7.58	1.49	**15.78**	8.92	4.87	2.53	1.27
	FPGD [25]	22.95	**26.54**	25.25	15.61	4.87	**16.28**	14.48	11.87	9.47	5.57
	N400 [26]	**57.59**	33.05	13.68	12.63	33.08	11.31	6.12	16.19	29.49	**37.11**
Fillers in [27]	SPR [27]	7.45	12.08	15.46	15.95	**16.05**	8.60	10.47	11.36	11.86	**13.33**
	FPGD [27]	8.98	8.05	8.45	9.04	**12.00**	8.94	10.91	12.91	13.81	**14.00**
	MAZE [28]	4.76	3.03	7.65	36.48	**84.83**	9.96	28.27	52.00	73.38	**88.64**
Michaelov+, 2024	N400 [29]	1.09	1.82	2.41	**2.12**	1.13	0.95	1.51	**1.70**	1.38	0.99
Federmeier+, 2007	N400 [30]	0.69	2.30	6.63	15.80	**18.84**	0.93	3.08	8.00	15.81	**18.89**
W&F, 2012	N400 [31]	0.36	0.18	0.10	0.36	**0.58**	0.41	0.20	0.14	0.25	**0.51**
Hubbard+, 2019	N400 [32]	0.20	0.21	0.27	**0.32**	0.24	0.13	0.14	0.25	**0.38**	0.34
S&F, 2022	N400 [33]	0.14	0.23	0.34	**0.50**	0.47	0.32	0.36	0.55	**0.62**	0.46
Szewczyk+, 2022	N400 [34]	1.09	2.45	4.26	**6.05**	5.84	2.00	3.41	4.92	6.50	**6.56**

に対して適用する Logit lens [35] を用いる：

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

ここで $\boldsymbol{W}_U \in \mathbb{R}^{d\times|V|}$ は埋め込み行列，LayerNorm$(\cdot)$ は最終層のレイヤー正規化，softmax$(\cdot)[\mathrm{id}(w)]$ は，語 w に付与された確率を取得していると読んでいただきたい．また Logit lens の発展手法として Tuned lens [36] も提案されてきた：

$$p(w|\boldsymbol{w}_{<t},\boldsymbol{h}_{l,t}) = \mathrm{LogitLens}((\boldsymbol{W}_l+\boldsymbol{1})\boldsymbol{h}_{l,t}+\boldsymbol{b}_l)[\mathrm{id}(w)]\ . \quad (2)$$

これは，logit-lens に追加して最終層の予測を近似できるようにアフィン変換 $\boldsymbol{W}_l, \boldsymbol{b}_l$ を各層で学習したものであり，レイヤ間の空間のずれにある程度対処する．実験節では，両手法で得られた結果を載せる．

2.3　読み活動・脳波データ

本研究では，Table 1 の各行に対応する 15 種の人間の読み活動・脳波データを用いる．このような認知モデリングで典型的に用いられる，自己ペース読み時間（SPR），視線停留時間（FPGD），脳波（N400），MAZE データを含めており，いずれにおいても，各単語 w_t に読み負荷 Cost(w_t) が付与されている．前処理は既存研究に従っている [7, 20, 28, 37]．また結果の解釈時に，人間の測定指標の違いとテキストの違いとの交絡を防ぐため，極力同じテキスト上で測られた異なる測定指標に基づくデータを含めた．

2.4　言語モデル

Logit-lens の実験では 21 モデル（付録 B）を，tuned-lens の実験では，そのうち学習済み tuned-lens パラメータ[2)]が公開されている 14 モデルを用いる．

3　実験結果

3.1　最終層が認知的に妥当とは限らない

図 1 に，例として DC，NS データ上で評価した層ごとの PPP を載せる．言語モデルの前半層で計算されるサプライザルの方が読み時間をうまく説明することがわかり，むしろ最終層を用いた場合は，認知的妥当性が過小評価されることが懸念される．表 1 に，より網羅的な結果を載せる．モデル内の相対的な層の位置ごとに，モデル横断的に平均した PPP を載せており（例えば 0-0.2 は，各モデルの前半 20%の層に対応），特に FPGD や SPR といった比較的速く，浅い処理を反映していると考えられる行動データは，前半層で高い PPP が得られている．[3)]

3.2　言語モデルのスケーリング

表 3 に，各データについて，モデルの大きさ（x 軸）と PPP（y 軸）の関係を 2 種類示している．黒縁のマーカと灰色の回帰直線は，各モデルの最終層か

2)　https://huggingface.co/spaces/AlignmentResearch/tuned-lens/tree/main
3)　検定は付録 D で行う．

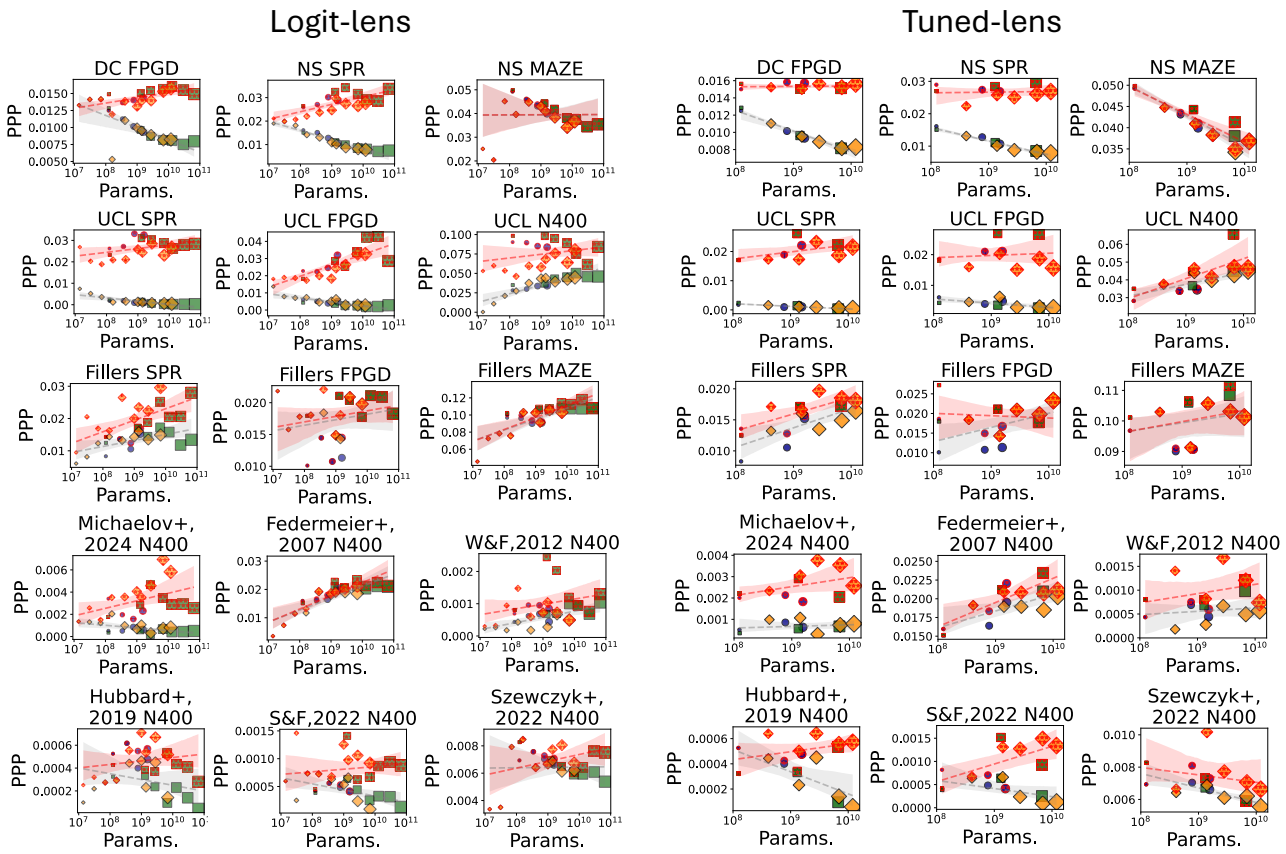

図 3　言語モデルの大きさと PPP の関係性．黒縁のマーカおよび灰色の回帰直線が最終層の PPP によるもの．赤縁のマーカおよび赤色の回帰直線が最も良い中間層の PPP によるもの．

ら得られた PPP に基づく結果であり，大きなモデルほど PPP は悪化する傾向にある既存の報告 [5, 6, 7] を再現している．赤縁のマーカと赤色の回帰直線は，各モデルの中間層から得られた最も良い PPP に基づく結果であり，最適なレイヤの選択のもとでは，より大きなモデルの方がより良い PPP を達成できることがわかる．すなわち，大きなモデルの中間層から得られる予測が人間らしい．

3.3　良い中間層は簡単に見つかる

ランダムに中間層を選んだ場合，どれぐらいの層が，最終層で達成されていたかつての PPP ベスト値を上回るだろうか．同モデルシリーズ（GPT-2, Llama2, Pythia）内で，最終層に基づいて得られていた最も良い PPP（しばしば小さなモデルから得られる）に対して，各中間層の PPP が勝るかを集計し，その勝率を報告する．勝率は各データ・モデルでおおむね 8 割を超え（詳細は付録中，表 2），たまたま外れ値的に良いレイヤがあるために，本研究において，大きい言語モデルが過大評価されているというわけではない．

3.4　いつ中間層が有効か？

中間層を活用することで，どのような単語において，読み活動予測誤差が減少しただろうか？最終層から、最も良い中間層のサプライザルに変更することで，残余誤差の減少が大きくなる単語を観察すると，長くて頻度の低い，いわゆる難易度の高い単語であった．この知見は，大規模言語モデルが，低頻度語で人間よりも驚かない（予測が正確すぎる）ために人間から逸脱して見えるという既存の知見と一致し [9]，中間層を活用することで、この問題が経験的に解消されていた．

4　おわりに

本研究では，言語モデルの中間層における次単語の予測で，人間の読み振る舞い・脳波がうまく説明できることを示した．言語モデルの各層がどのような言語情報を有するかといった，解釈性方面の知見と統合することで，人間と言語モデルの文処理について洞察が得られることを期待する．

謝辞

MBZUAI の Dr. Yova Kementchedjhieva，Dr. Ted Broscoe との議論に感謝する．本研究は，JST さきがけ JPMJPR21C2，JSPS 科研費 24H00087，および JST CREST JPMJCR20D2 の支援を受けたものです．

参考文献

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

[2] Roger Levy. Expectation-based syntactic comprehension. **Journal of Cognition**, Vol. 106, No. 3, pp. 1126–1177, 2008.

[3] Andy Clark. Whatever next? predictive brains, situated agents, and the future of cognitive science. **Behav. Brain Sci.**, Vol. 36, No. 3, pp. 181–204, June 2013.

[4] Nathaniel J Smith and Roger Levy. The effect of word predictability on reading time is logarithmic. **Cognition**, Vol. 128, No. 3, pp. 302–319, September 2013.

[5] Byung-Doh Oh and William Schuler. Why does surprisal from larger transformer-based language models provide a poorer fit to human reading times? **TACL**, Vol. 11, pp. 336–350, March 2023.

[6] Cory Shain, Clara Meister, Tiago Pimentel, Ryan Cotterell, and Roger Philip Levy. Large-scale evidence for logarithmic effects of word predictability on reading time. **PsyArXiv**, 2022.

[7] Tatsuki Kuribayashi, Yohei Oseki, and Timothy Baldwin. Psychometric predictive power of large language models. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, **Findings of the Association for Computational Linguistics: NAACL 2024**, pp. 1983–2005, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

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

[9] Byung-Doh Oh, Shisen Yue, and William Schuler. Frequency explains the inverse correlation of large language models' size, training data amount, and surprisal's fit to reading times. In Yvette Graham and Matthew Purver, editors, **Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 2644–2663, St. Julian's, Malta, March 2024. Association for Computational Linguistics.

[10] Byung-Doh Oh and William Schuler. Leading whitespaces of language models' subword vocabulary pose a confound for calculating word probabilities. In **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing**, pp. 3464–3472, November 2024.

[11] Mario Giulianelli, Luca Malagutti, Juan Luis Gastaldi, Brian DuSell, Tim Vieira, and Ryan Cotterell. On the proper treatment of tokenization in psycholinguistics. In **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing**, pp. 18556–18572, November 2024.

[12] Adam Goodkind and Klinton Bicknell. Predictive power of word surprisal for reading times is a linear function of language model quality. In **Proceedings of CMCL**, pp. 10–18, 2018.

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

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

[15] Tiago Pimentel, Clara Meister, Ethan G. Wilcox, Roger Levy, and Ryan Cotterell. On the effect of anticipation on reading times. **arXiv preprint**, 2022.

[16] Olaf Dimigen, Werner Sommer, Annette Hohlfeld, Arthur M Jacobs, and Reinhold Kliegl. Coregistration of eye movements and EEG in natural reading: analyses and review. **J. Exp. Psychol. Gen.**, Vol. 140, No. 4, pp. 552–572, November 2011.

[17] Keith Rayner and Charles Clifton, Jr. Language processing in reading and speech perception is fast and incremental: implications for event-related potential research. **Biol. Psychol.**, Vol. 80, No. 1, pp. 4–9, January 2009.

[18] Ellen F Lau, Colin Phillips, and David Poeppel. A cortical network for semantics: (de)constructing the N400. **Nat. Rev. Neurosci.**, Vol. 9, No. 12, pp. 920–933, December 2008.

[19] Naoko Witzel, Jeffrey Witzel, and Kenneth Forster. Comparisons of online reading paradigms: eye tracking, moving-window, and maze. **J. Psycholinguist. Res.**, Vol. 41, No. 2, pp. 105–128, April 2012.

[20] Andrea Gregor de Varda, Marco Marelli, and Simona Amenta. Cloze probability, predictability ratings, and computational estimates for 205 english sentences, aligned with existing EEG and reading time data. **Behav. Res. Methods**, Vol. 56, No. 5, pp. 5190–5213, August 2024.

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

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

[23] Richard Futrell, Edward Gibson, Harry J Tily, Idan Blank, Anastasia Vishnevetsky, Steven T Piantadosi, and Evelina Fedorenko. The natural stories corpus: a reading-time corpus of english texts containing rare syntactic constructions. **Lang. Resour. Eval.**, Vol. 55, No. 1, pp. 63–77, 2021.

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

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

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

[27] Shravan Vasishth, Katja Suckow, Richard L Lewis, and Sabine Kern. Short-term forgetting in sentence comprehension: Crosslinguistic evidence from verb-final structures. **Lang. Cogn. Process.**, Vol. 25, No. 4, pp. 533–567, May 2010.

[28] Michael Hahn, Richard Futrell, Roger Levy, and Edward Gibson. A resource-rational model of human processing of recursive linguistic structure. **Proc. Natl. Acad. Sci. U. S. A.**, Vol. 119, No. 43, p. e2122602119, October 2022.

[29] James A. Michaelov, Megan D. Bardolph, Cyma K. Van Petten, Benjamin K. Bergen, and Seana Coulson. Strong Prediction: Language Model Surprisal Explains Multiple N400 Effects. **Neurobiology of Language**, Vol. 5, No. 1, pp. 107–135, 04 2024.

[30] Kara D Federmeier, Edward W Wlotko, Esmeralda De Ochoa-Dewald, and Marta Kutas. Multiple effects of sentential constraint on word processing. **Brain Res.**, Vol. 1146, pp. 75–84, May 2007.

[31] Edward W Wlotko and Kara D Federmeier. So that's what you meant! event-related potentials reveal multiple aspects of context use during construction of message-level meaning. **Neuroimage**, Vol. 62, No. 1, pp. 356–366, August 2012.

[32] Ryan J Hubbard, Joost Rommers, Cassandra L Jacobs, and Kara D Federmeier. Downstream behavioral and electrophysiological consequences of word prediction on recognition memory. **Front. Hum. Neurosci.**, Vol. 13, p. 291, August 2019.

[33] Jakub M Szewczyk and Kara D Federmeier. Context-based facilitation of semantic access follows both logarithmic and linear functions of stimulus probability. **J. Mem. Lang.**, Vol. 123, No. 104311, p. 104311, April 2022.

[34] Jakub M Szewczyk, Emily N Mech, and Kara D Federmeier. The power of "good": Can adjectives rapidly decrease as well as increase the availability of the upcoming noun? **J. Exp. Psychol. Learn. Mem. Cogn.**, Vol. 48, No. 6, pp. 856–875, June 2022.

[35] nostalgebraist. interpreting gpt: the logit lens., 2020.

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

[37] James Michaelov, Catherine Arnett, and Ben Bergen. Revenge of the fallen? recurrent models match transformers at predicting human language comprehension metrics. In **First Conference on Language Modeling**, 2024.

[38] Keith Rayner. Eye movements in reading and information processing: 20 years of research. **Psychol. Bull.**, Vol. 124, No. 3, pp. 372–422, 1998.

[39] Robyn Speer. rspeer/wordfreq: v3.0, September 2022.

[40] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. Vol. 1, No. 8, p. 9, 2019.

[41] 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. OPT: Open Pre-trained Transformer Language Models. **arXiv preprint**, Vol. cs.CL/2205.01068v4, , 2022.

[42] Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, KyleO'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff ほか. Pythia: A suite for analyzing large language models across training and scaling. In **International Conference on Machine Learning**, pp. 2397–2430. PMLR, 2023.

[43] Kenneth I Forster, Christine Guerrera, and Lisa Elliot. The maze task: measuring forced incremental sentence processing time. **Behav. Res. Methods**, Vol. 41, No. 1, pp. 163–171, February 2009.

表 2　同モデルシリーズの最終層で得られていた PPP ベスト値に対する，中間層の PPP の勝率．勝率が 1 に近い場合，どの中間層を適当に選んでも，かつてのベスト PPP 値に勝っていることを意味する．“PT” は “Pythia” である．本表では，層数が多い 1B 以上のモデルと，最終層と中間層で比較的結果が大きく変わった読み活動データに着目している．

	Logit-lens (win rate)												Tuned-lens (win rate)						
Data	GPT2 XL	OPT 1.3B	OPT 2.7B	OPT 6.7B	OPT 13B	OPT 30B	OPT 66B	PT 1B	PT 1.4B	PT 2.8B	PT 6.9B	PT 12B	GPT2 XL	OPT 1.3B	OPT 6.7B	PT 1.4B	PT 2.8B	PT 6.9B	PT 12B
DC FPGD [22]	0.80	0.80	0.82	0.76	0.73	0.76	0.78	0.00	0.32	0.36	0.73	0.73	0.73	0.80	0.67	0.64	0.58	0.55	0.54
NS SPR [23]	0.82	0.80	0.85	0.76	0.73	0.76	0.85	0.47	0.52	0.45	0.21	0.41	0.55	0.76	0.70	0.56	0.39	0.36	0.41
UCL SPR [25]	0.78	0.80	0.79	0.76	0.76	0.78	0.80	0.71	0.52	0.64	0.70	0.70	0.73	0.80	0.61	0.52	0.36	0.42	0.35
UCL FPGD [25]	0.94	0.88	0.82	0.79	0.83	0.88	0.83	0.59	0.56	0.58	0.70	0.73	0.90	0.92	0.91	0.96	0.48	0.73	0.73

表 3　UCL における，人間の他の計測指標に対する結果．

		Logit-lens (PPP)					Tuned-lens (PPP)				
テキスト	指標	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
UCL	SPR [25]	24.51	23.35	18.84	7.80	1.81	15.78	8.92	4.87	2.53	1.27
	FPGD [25]	22.62	26.24	25.12	15.55	5.02	16.28	14.48	11.87	9.47	5.57
	SPGD [25]	241.76	191.77	124.82	30.13	3.97	68.43	20.05	5.28	1.71	3.87
	ELAN (125–175ms) [26]	0.79	0.38	0.18	0.30	0.49	0.22	0.38	0.77	0.67	0.50
	LAN (300–400ms) [26]	70.32	48.69	25.78	5.05	6.14	19.94	2.74	1.72	5.02	8.76
	N400 (300–500ms) [26]	57.45	33.30	14.01	12.89	32.26	11.31	6.12	16.19	29.49	37.11
	EPNP (400–600ms) [26]	78.22	58.65	35.25	7.76	3.09	29.31	6.67	1.29	1.56	4.45
	P600 (500–700ms) [26]	69.52	50.43	29.25	6.37	7.39	17.70	3.30	1.80	5.83	11.36
	PNP (600–700ms) [26]	60.47	46.76	29.58	7.32	1.56	25.38	8.09	2.30	0.73	1.61

A　回帰モデル

以下の回帰モデルを用いて，読み活動データのモデリングを行った：

$$\begin{aligned}\text{Cost}(w_t) \sim\ & \text{surprisal}(w_t) + \text{surprisal}(w_{t-1}) + \text{surprisal}(w_{t-2}) + \text{単語長}\ (w_t) + \text{頻度}\ (w_t) \\ & + \text{単語長}\ (w_{t-1}) + \text{頻度}\ (w_{t-1})\ \text{単語長}\ (w_{t-2}) + \text{頻度}\ (w_{t-2})\ .\end{aligned}$$

なお，直近の単語の処理負荷が影響を与えるスピルオーバー効果 [38] に対処するため，$\text{surprisal}(w_{t-1})$ と $\text{surprisal}(w_{t-2})$ をベースライン素性に常に含めている．単語長は文字数で数え，頻度は word freq パッケージで求め [39]，サプライザルに倣い対数へ変換している．また N400 データについて，Michaelov+24 [29] のデータのみ，各電極の情報が分けて記録されており，ランダム効果として電極位置（番号）も回帰モデルに含んでいる．

B　言語モデル

21 モデルの内訳は以下の通りである：GPT-2 (124M, 355M, 774M, and 1.5B parameters) [40], OPT (125M, 1.3B, 2.7B, 6.7B, 13B, 30B, and 66B parameters) [41], and Pythia (14M, 31M, 70M, 160M, 410M, 1B, 1.4B, 2.8B, 6.9B, and 12B parameters) [42]. Tuned-lens 実験については，次の 14 モデルを用いている：GPT-2 124M, 774M, and 1.5B; OPT 125M, 1.3B, and 6.7B; and Pythia 70M, 160M, 410M, 1B, 1.4B, 2.8B, 6.9B, and 12B.

C　人間の読み活動・脳波データ

FPGD は視線計測により得られる指標であり，ある単語に最初に視線が停留してから初めて別の単語に視線が出るまでの時間を表す．SPR は固定スライディング窓を移動させながら（典型的には単語ごと）文章を読んだときに，各単語を見た時間を指す．視線計測に対する安価な代替手段として用いられることが多い．N400 はある単語が見せられてから，およそ 400ms 後に生じる事象関連電位であり，前処理は既存研究に従っている [37]．MAZE では，単語レベルの SPR と同様に，1 単語ずつ提示して被験者に文を読ませる．ただし，各単語は，これまでの文脈の続きとして不適切な単語とともに表示され，どちらが続きとして自然な候補かという 2 択問題を解きながら文を読む．このときの問題の回答速度を読み時間として付与している [43, 24].

D　有効な層と計測指標間の関係

各実験設定で得られた PPP を収集し，これら PPP を実験設定から予測する回帰モデルを考える．

$$\text{PPP}(s) \sim \text{stimuli}(s) + \text{model}(s) + \text{layer\_depth}(s) + \text{measure}(s) + \text{lens}(s) + \text{layer\_depth}(s) : \text{measure}(s)\ ,$$

stimuli はデータセットで用いられているテキスト（表 1 中「テキスト」列）を，model は用いた言語モデル，layer_depth は言語モデルのレイヤ位置，measure は人間の測定指標（表 1 中「指標」列），$\text{lens}(s)$ は用いた手法（logit-lens/tuned-lens）をエンコードしている．我々の関心は相互作用項 $\text{layer\_depth}(s) : \text{measure}(s)$ にあり，measure に応じて PPP の高さを説明する layer_depth が変わるかが知りたい．なお，measure はカテゴリカルな変数であり，FPGD をダミークラスとしている．回帰モデルのフィット後，相互作用項の係数を確認すると，layer_depth : SPR の係数は有意に 0 より小さく，layer_depth : N400 と layer_depth : MAZE は有意に 0 より大きかった（1 サンプル t-test）．したがって，測定指標に応じて，高い PPP に紐づく層は異なる傾向にあると確認され，特に遅い反応である N400 と MAZE は後半層に紐づくことが分かった．

なお，UCL 上ではより多くの計測指標についても実験を行っており，比較的速く，浅い処理を反映していると想定される反応（FPGD, SPR）は前半層，比較的遅い反応（N400, MAZE）は後半層という対応について，FPRD，SPR，N400，MAZE 以外の指標では必ずしも成り立っていなかった．したがって，あくまでこういった傾向は今回分析した 4 つの指標について言えることである．これら 4 つの指標は，予測に基づく文処理で典型的に扱われる指標であり，逆に言えば，その他の SPGD (second pass gaze duration) や，ELAN，P600 といった他の指標が，そもそもサプライザルで説明されるべき処理を反映しているのかといった疑問もある．