大規模言語モデルの文処理は人間らしいのか？

栗林 樹生 [1]　大関 洋平 [2]　Timothy Baldwin[1,3]
[1]MBZUAI　[2] 東京大学　[3] メルボルン大学
{tatsuki.kuribayashi, timothy.baldwin}@mbzuai.ac.ae
oseki@g.ecc.u-tokyo.ac.jp

概要

言語モデルの計算する次単語確率は，人間の読み活動（読み時間など）をうまく説明する．この知見のもと，本研究では，教師あり指示学習（インストラクション・チューニング）済みの大規模言語モデルが，通常の大規模言語モデルと比べて，人間の読み活動と乖離した単語確率を計算する傾向を示す．すなわち，人間に好まれる応答をするよう言語モデルを追調整することは，心理言語学的な意味でモデルを人間に近づけることを含意していない．さらに，読み時間の予測に有効な指示（プロンプト）を調査する．特定の言語学的仮説を言語化した指示が有効であったものの，依然として純粋な次単語確率の方が人間の読み振る舞いを予測できた．

1　はじめに

言語モデルの計算する次単語確率が人間の読み活動（読み時間など）をうまく再現できる [1, 2, 3]．このことから，人間も続く単語の予測を行いながら文を読んでいることが支持されている [4, 5]．この知見を踏まえ計算心理言語学領域では，人間の文処理の解明を目指し，どのような言語モデル・アルゴリズムの計算する確率が，人間の振る舞いをより忠実に再現できるのか探求されている（図 1；[6, 7, 8]）．

一方自然言語処理分野では，人間に好まれる返答を出力するよう訓練された大規模言語モデル（教師あり指示学習済み大規模言語モデル）が成功を収めている．このような発展を踏まえると，**人間に好まれるように追調整した大規模言語モデルが，人間の読み活動の再現という文脈でも優位性をもつのか**が問いとして浮上し，その答えには複数の可能性が考えられる．例えば，人間に好まれる出力を促すという学習は，広い意味では人間と対応をとる方向にあり（例えば，人間も言語モデルも文章を読んでいるときに，でっちあげられた情報を期待しない/すべきでないだろう [9, 10]），結果的に人間らしい振る舞いを促すかもしれない．一方で，教師あり学習を経た言語モデルは，もはや大規模コーパスの統計情報を忠実に計算する狭義の言語モデルではないだろう．仮に人間の文処理の本質が大量の言語刺激の集積に基づく次単語の予測にあるとするならば，特定のデータに対する教師あり指示学習は，この文脈では不必要かもしれない．さらに，教師あり指示学習で目指しているものは，心理言語学が典型的に対象とする一人間の能力のモデルではなく，どのような話題にも適切に回答できるいわば超知能的な存在であろう．この観点では，近年の教師あり指示学習が，心理言語学的な意味での人間らしさをモデルに促すと期待はできないかもしれない．

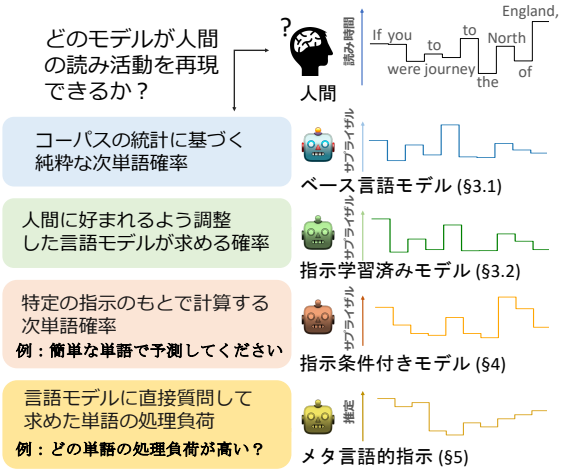

図 1　人間と言語モデルの「読み滞り」を比較する．人間の読み時間と言語モデルの計算するサプライザル（$-\log p$(単語 | 文脈)）を対比させ，(i) 通常の大規模言語モデル，(ii) 指示学習済み大規模言語モデル，(iii) 指示条件付き大規模言語モデル，(iv) メタ言語的指示による読み時間推定，のいずれがより正確に人間の読み活動を再現するのか調べる．

実験から，指示学習済み大規模言語モデルは，同程度の正確性（パープレキシティ，PPL）をもつ指示学習なしの言語モデルと比較して，**人間の言語処**

理をうまく再現できないことがわかった．すなわち現在の指示学習の成功は，認知モデリングを進展させていない．

さらに，指示学習と認知モデリングという新たな組み合わせから，(i) プロンプトの選択によって言語モデルが人間の振る舞いをよりうまく再現できるようになるのか，(ii) モデルに直接逐次的文処理の負荷を問い合わせることで（メタ言語的指示），確率に基づく推定よりも逐次的処理負荷をうまく推定できるのかといった問いに答える．前者の問いについては，言語学的に動機づけられた特定の指示が人間らしい振る舞いを促したものの，依然として通常の言語モデルの計算する確率の方が人間の読み振る舞いを予測できた．後者の問いについては，メタ言語的指示に基づく読み負荷の推定は，純粋な確率に基づく推定に対して劣ることがわかった．まとめると，近年の指示学習済み言語モデルの進展は，人間の読み活動の再現において，純粋な次単語確率に勝る道具立てを提供できていないことがわかる．またこのことは，計算言語学の発展における，単語の確率値を提供できる「開かれた API」の価値も示している．

2 人間の文処理のモデリング

2.1 橋渡し仮説

人間の読み時間（RT）などで測定される文の逐次的な処理負荷は，言語モデル θ によって推定される単語のサプライザル $h_{t,\theta}(w)$ で説明される：

$$\mathrm{RT}(w_t) \sim h_{t,\theta}(w_t) + \mathrm{baselines}(w_t) \ , \tag{1}$$

$$h_{t,\theta}(w) := -\log_2 p_\theta(w|\boldsymbol{w}_{<t}) \ . \tag{2}$$

本実験では，読み時間の予測に対するサプライザル因子の寄与が焦点である．具体的には，サプライザル因子とその他ベースライン素性（単語の長さなど）を用いた回帰モデルが，ベースライン素性のみのモデルに対して，どの程度読み時間予測能力が向上するのかを測る．この予測能力の差を PPP（psychometric predictive power）と呼び，この値が高いほど，追加したサプライザル因子が読み時間の説明に寄与している．そして，この論文ではどのような言語モデル θ でサプライザルを計算することで，高い PPP が得られるのかを調査する．なお，既存研究に基づき [11, 12, 13]，サプライザルの期待値であるシャノンエントロピ $\mathrm{H}_\theta(W_t)$ [14] と，その一般化されたものであるレニエントロピ $\mathrm{H}_{\alpha,\theta}(W_t)$ [15] を用いた場合の PPP も報告する：[1)]

$$\mathrm{H}_\theta(W_t) := \mathop{\mathbb{E}}_{w\sim p(\cdot|\boldsymbol{w}_{<t})} h_{t,\theta}(w) \ , \tag{3}$$

$$\mathrm{H}_{\alpha,\theta}(W_t) := \lim_{\gamma\to\alpha} \frac{1}{1-\gamma} \log_2 \sum_{w\in W} p_\theta(w|\boldsymbol{w}_{<t})^\gamma \ . \tag{4}$$

$\alpha = 1$ の場合に式 3 と式 4 は一致し，既存研究 [13, 16] で有効性が示されている $\alpha = 0.5$ の場合と共に，結果を報告する．

2.2 実験設定

モデル： モデルの大きさや教師あり指示学習の有無が異なる 26 モデル（OPT, GPT-2/3/3.5, Falcon LM, LLaMA-2）を対象とした（付録参照）．[2)]

データ： Dundee Corpus (DC) [17] と Natural Stories Corpus (NS) [18] を用いる．紙面スペースの都合上，図 2 から NS の結果は省いているが，一貫した結果が得られている．データの前処理については，付録を参照されたい．

3 実験 1: 大規模言語モデル

§3.1 では，教師あり指示学習をしていない大規模言語モデル（ベース言語モデル）の PPP を調査する．§3.2 では，指示学習済みモデルを分析し，§4 と §5 では，さらに指示方式を調査する．

3.1 ベース言語モデル

図 2 左に，大規模言語モデルの PPP と PPL の関係を示す[3)]．各図形が各モデルに対応し，ここでは枠線のないベース言語モデルと，それらから求めた PPL–PPP 回帰直線に注目する．まず，すべてのモデルにおいて PPP は統計的に有意に正であり，既存研究の通り [19]，サプライザル仮説は支持された．次に，レニエントロピ（$\alpha = 0.5$）を用いた際の PPP は，サプライザルやシャノンエントロピを用いた場合よりも高く，既存研究 [13, 16] よりも幅広いモデルでその有効性が示された．さらに，事前学習済み言語モデルにおいて，PPL の低い言語モデルほど PPP が悪化するというある種の逆スケーリング則が報告されており [8, 19, 20, 21]，回帰直線の傾きからこの知見も再現された．

1) 語彙集合 W は，言語モデルの語彙語彙で近似する [13].
2) GPT-3/3.5 については，API の仕様上 top-k 単語の確率しか取得できないため，エントロピの結果は示していない．また，GPT-4 は確率値が取得できないことから除外している．
3) PPL はいずれのモデルもコーパス上で読み時間が付与されている単位に対する平均的な確率に基づいて計算しており，トークナイザの違いによる影響は受けない

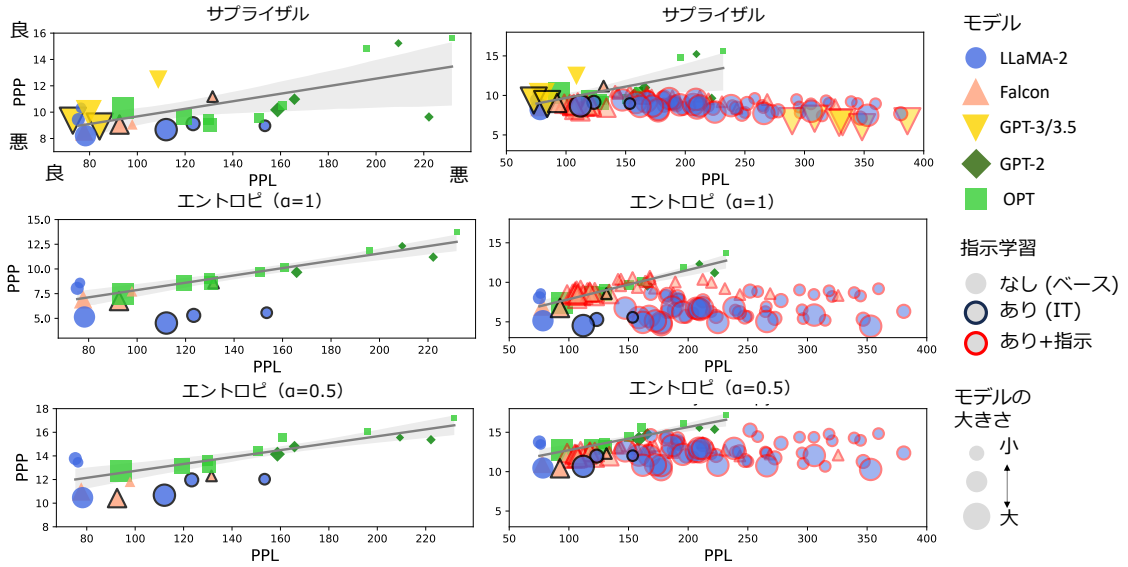

図 2　PPL と PPP の関係．各図形は各言語モデルや設定（指示）に対応する．回帰直線はベース言語モデルの結果から求めており，灰色の領域は 95%信頼区間である．左図はベース言語モデルと指示学習済みモデルの結果のみを示しており，右図はさらに指示条件付き言語モデルの結果を重ね合わせたものである．

3.2　指示学習済み言語モデル

図 2 左より，指示学習済みモデル（黒枠線）の多くが，ベース言語モデルによって推定された PPL–PPP 回帰直線に対して右下側に位置していることがわかる．指示学習済みモデルは PPL が悪化する傾向にあり，それにも関わらず PPP が向上していない．すなわち，**指示学習済みモデルは PPL と PPP の両立ができておらず**，コーパスのモデリングという観点でも，人間の認知モデリングという観点でも，これらは質の悪い（狭義の）言語モデルであると言える．

4　実験 2: 指示条件付き言語モデル

特定の指示のもとで次の単語を予測させることで，サプライザルの推定に言語的なバイアスがかけられる．この観点で有効な指示を探し，人間特有の文処理のバイアスの言語化を狙う．

本稿では，表 1 に示す 8 つの指示を試す．特定の言語的側面（語彙選択・文法性）に対するバイアス（単純・複雑にする等）の導入を指示している．また，「次の文を補完してください」という指示と，指示をしない場合（§3.2 と同様）も合わせて報告する．

指示の影響：　表 1 に，それぞれの指示を用いた場合の PPP を示す．なお以降の節では，指示学習済みの言語モデル（7 種類）による結果の平均を示す．結果として，(i) 特定の指示では PPP が向上すること，(ii) 文法に言及すると PPP が向上すること，また (iii) 全体的な傾向としては「単純な語・文法で補完してください」といった指示で，人間の読み活動に近いサプライザルが計算された．このような**単純バイアス**は，人間の文処理において，不必要に複雑な解析をしないという good-enough 処理仮説 [22] と接続ができ，間接的にそのような仮説を支持していると解釈できる．

PPP と PPL の関係：　図 2 左に対し，指示条件付き言語モデルの結果を重ね合わせたものが図 2 右である．各図形は，特定の指示学習モデルに対して特定の支持を与えた際の結果である．特定の指示で PPP は改善するものの，依然として指示条件付きモデルは，ベース言語モデルが設定する PPP と PPL の回帰直線の右下に位置し，やはり PPP と PPL の両立という観点では，ベース言語モデルの計算する単なる次単語確率が優れていると言える．

5　実験 3: メタ言語的指示

ここまで，言語モデルが計算する確率によって処理負荷を説明しようとしてきたが，指示学習済みモデルであれば，「この単語の読み時間はどの程度か」といった質問に直接答えてくれるだろう．このような**メタ言語的指示** [23] と従来の確率に基づく読み時間の再現では，どちらが優れているのかを評価する．

設定：　事前分析より「文中の各単語の読み時間を答えてください」といった類の質問では，同じ数

表 1　異なる指示を用いたときの PPP の変化．プロンプト有り設定について最も値の大きい PPP を**太字**にしている．

番号	プロンプト	DC			NS		
		$h\uparrow$	H $\uparrow$	$H_{0.5}\uparrow$	$h\uparrow$	H $\uparrow$	$H_{0.5}\uparrow$
1	Please complete the following sentence to make it as **grammatically simple** as possible:\n $w_0, \cdots, w_{t-1}$	8.23	7.46	12.26	**6.55**	2.62	8.26
2	Please complete the following sentence with a careful **focus on grammar**: \n $w_0, \cdots, w_{t-1}$	**8.24**	7.19	11.99	6.20	2.99	8.72
3	Please complete the following sentence to make it as **grammatically complex** as possible: \n $w_0, \cdots, w_{t-1}$	7.77	6.99	11.74	5.66	2.54	7.75
4	Please complete the following sentence using the **simplest vocabulary** possible: \n $w_0, \cdots, w_{t-1}$	7.82	**7.48**	12.15	5.70	**3.11**	**8.90**
5	Please complete the following sentence with a careful **focus on word choice**: \n $w_0, \cdots, w_{t-1}$	7.87	6.86	11.50	6.06	2.94	8.60
6	Please complete the following sentence using the most **difficult vocabulary** possible: \n $w_0, \cdots, w_{t-1}$	7.31	6.71	11.38	4.73	2.43	7.57
7	Please complete the following sentence **in a human-like manner**. It has been reported that human ability to predict next words is weaker than language models and that humans often make noisy predictions, such as careless grammatical errors.\n $w_0, \cdots, w_{t-1}$	7.86	7.30	12.34	4.60	3.03	8.78
8	Please complete the following sentence. We are trying to **reproduce human reading times with the word prediction probabilities you calculate**, so please predict the next word like a human. It has been reported that human ability to predict next words is weaker than language models and that humans often make noisy predictions, such as careless grammatical errors.\n $w_0, \cdots, w_{t-1}$	8.17	7.36	**12.42**	4.83	**3.11**	8.73
9	Please complete the following sentence: \n $w_0, \cdots, w_{t-1}$	8.34	7.12	11.88	5.77	3.01	8.74
10	w/o prompting	9.32	6.15	11.48	6.25	2.69	8.86

値を繰り返し答えるなど望んだ回答が得られなかった．これを踏まえて，「この文の中の単語を処理負荷・サプライザルが高い順に並べてください」という質問に問題を簡略化させた（表 2）．異なる読み時間コーパスから抽出した文と正しい単語の並びの例を 3 つ見せ（付録参照），コーパス中の各文について処理負荷順に単語を並び替える問題を解かせる．文ごとに正しい並びとモデルが出力した並びの間のスピアマン順位相関を求め，そのコーパス平均を報告する．サプライザルの高い順に単語を並べるベースラインと比べて，メタ言語的指示による並び替えの相関が高いかが焦点である．

結果：　表 2 に結果を示す．サプライザルに基づく並び替えよりも，メタ言語的指示で得られた並び替え結果のほうが相関が低く，サプライザルに基づく認知モデリングの有効性がここでも示された．[4] なお，実際のサプライザルに基づく並び替えと，サプライザルを間接的に求めさせる表 2 中 2 つ目の指示で得られた並び替えの間の相関は 0.1–0.2 程度であり，文に対する自身の「驚き」に関するある種のメタ認知ができていないことも示唆された．

4）　多くの要素を並び替えて順に出力していくこと自体がモデルにとって難しいという懸念を踏まえ，モデルが出力した最初の 5 単語で相関を計算したが，その場合もやはり無相関に近い値であった．

表 2　推定された読み負荷と実際の値の順位相関．メタ言語的指示の結果は異なる例示による 3 回の実験の平均値．

方法論（プロンプト）	モデル	DC $\uparrow$	NS $\uparrow$
Suppose humans read the following sentence: [SENT]. List the tokens in order of their reading cost (high to low) during sentence processing.	LLaMA-2 7B	0.09	-0.04
	LLaMA-2 13B	0.06	-0.03
	Falcon 7B	0.12	0.01
	Falcon 40B	0.03	-0.03
	GPT3.5 D2	0.05	0.05
	GPT3.5 D3	0.08	0.03
Suppose you read the following sentence: [SENT]. List the tokens in order of their probability in context (low to high).	LLaMA-2 7B	0.05	0.00
	LLaMA-2 13B	0.04	0.06
	Falcon 7B	0.08	0.05
	Falcon 40B	0.02	0.13
	GPT3.5 D2	0.03	0.02
	GPT3.5 D3	-0.01	0.06
サプライザルによる並び替え	LLaMa-2 7B	0.28	0.19
	LLaMa-2 13B	0.27	0.19
	Falcon 7B	0.32	0.18
	Falcon 40B	0.28	0.17
	GPT3.5 D2	0.28	0.16
	GPT3.5 D3	0.25	0.17

6　おわりに

本研究では，認知モデリングの視点から，大規模言語モデルに対する教師あり指示学習など，自然言語処理における新しい概念の位置づけを調査した．人間と言語モデルの対応付け（AI アライメント）に対して分野の関心が高まっているが，認知モデリングもまた計算モデルを人間に対応付ける試みであり，その出口は工学的・科学的な意味で異なるが，両領域が示唆を与え合うことに期待したい．

謝辞

本研究は、JST さきがけ JPMJPR21C2 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] 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.

[3] 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.

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

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

[6] Ethan Gotlieb Wilcox, Jon Gauthier, Jennifer Hu, Peng Qian, and Roger Levy. On the Predictive Power of Neural Language Models for Human Real-Time Comprehension Behavior. In **Proceedings of CogSci**, pp. 1707–1713, 2020.

[7] Byung-Doh Oh, Christian Clark, and William Schuler. Surprisal estimators for human reading times need character models. In **Proceedings of ACL-IJCNLP 2021**, pp. 3746–3757, August 2021.

[8] 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.

[9] Herbert P Grice. Logic and conversation. In **Speech acts**, pp. 41–58. Brill, 1975.

[10] Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, et al. A general language assistant as a laboratory for alignment. **arXiv preprint arXiv:2112.00861**, 2021.

[11] Brian Roark, Asaf Bachrach, Carlos Cardenas, and Christophe Pallier. Deriving lexical and syntactic expectation-based measures for psycholinguistic modeling via incremental top-down parsing. In **Proceedings of EMNLP 2009**, pp. 324–333, August 2009.

[12] Marten van Schijndel and Tal Linzen. Can Entropy Explain Successor Surprisal Effects in Reading? In **Proceedings of SCiL 2019**, pp. 1–7, 2019.

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

[14] C E Shannon. A Mathematical Theory of Communication. **Bell System Technical Journal**, Vol. 27, No. 3, pp. 379–423, 1948.

[15] Alfréd Rényi. On measures of entropy and information. In **Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics**, Vol. 4.1, pp. 547–562. University of California Press, January 1961.

[16] Tong Liu, Iza Škrjanec, and Vera Demberg. Improving fit to human reading times via temperature-scaled surprisal, November 2023.

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

[18] Richard Futrell, Edward Gibson, Harry J. Tily, Idan Blank, Anastasia Vishnevetsky, Steven Piantadosi, and Evelina Fedorenko. The natural stories corpus. In **Proceedings of LREC 2018**, pp. 76–82, May 2018.

[19] 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.

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

[21] 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.

[22] Fernanda Ferreira and Matthew W Lowder. Chapter six - prediction, information structure, and Good-Enough language processing. In Brian H Ross, editor, **Psychology of Learning and Motivation**, Vol. 65, pp. 217–247. Academic Press, January 2016.

[23] Jennifer Hu and Roger Levy. Prompting is not a substitute for probability measurements in large language models. In **Proceedings of EMNLP 2023 (to appear)**, 2023.

[24] Merity Stephen, Xiong Caiming, Bradbury James, Socher Richard, et al. Pointer sentinel mixture models. In **Proceedings of ICLR 2017**, 2017.

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

表 3　使用したモデル

モデル	指示学習	リンク	量子化
GPT-2 117M		https://huggingface.co/gpt2	
GPT-2 355M		https://huggingface.co/gpt2-medium	
GPT-2 774M		https://huggingface.co/gpt2-large	
GPT-2 1.5B		https://huggingface.co/gpt2-xl	
LLaMa2 7B		https://huggingface.co/meta-llama/Llama2-7b-hf	
LLaMa2 7B	✓	https://huggingface.co/meta-llama/Llama2-7b-chat-hf	
LLaMa2 13B		https://huggingface.co/meta-llama/Llama2-13b-hf	8bits
LLaMa2 13B	✓	https://huggingface.co/meta-llama/Llama2-13b-chat-hf	8bits
LLaMa2 70B		https://huggingface.co/meta-llama/Llama2-70b-hf	4bits
LLaMa2 70B	✓	https://huggingface.co/meta-llama/Llama2-70b-chat-hf	4bits
Falcon 7B		ttps://huggingface.co/tiiuae/falcon-7b	
Falcon 7B	✓	https://huggingface.co/tiiuae/falcon-7b-instruct	
Falcon 40B		https://huggingface.co/tiiuae/falcon-40b	4bits
Falcon 40B	✓	https://huggingface.co/tiiuae/falcon-40b-instruct	4bits
OPT 125M		https://huggingface.co/facebook/opt-125m	
OPT 350M		https://huggingface.co/facebook/opt-350m	
OPT 1.3B		https://huggingface.co/facebook/opt-1.3b	
OPT 2.7B		https://huggingface.co/facebook/opt-2.7b	
OPT 6.7B		https://huggingface.co/facebook/opt-6.7b	
OPT 13B		https://huggingface.co/facebook/opt-13b	
OPT 30B		https://huggingface.co/facebook/opt-30b	
OPT 66B		https://huggingface.co/facebook/opt-66b	
GPT-3 babbage-002		accessed on 2023/10/20 for §3, and on 2023/11/04 for §4 and §5	
GPT-3 davinci-002		accessed on 2023/10/20 for §3, and on 2023/11/04 for §4 and §5	
GPT-3.5 text-davinci-003	✓	accessed on 2023/10/20 for §3, and on 2023/11/04 for §4 and §5	
GPT-3.5 text-davinci-002	✓	accessed on 2023/10/20 for §3, and on 2023/11/04 for §4 and §5	

A　回帰式

以下の回帰式を用いた：$\texttt{time}(w_t) \sim \texttt{surprisal}(w_t) + \texttt{surprisal}(w_{t-1}) + \texttt{surprisal}(w_{t-2}) + \texttt{length}(w) + \texttt{freq}(w) + \texttt{length}(w_{t-1}) + \texttt{freq}(w_{t-1}) + \texttt{length}(w_{t-2}) + \texttt{freq}(w_{t-2})$．サプライザル因子 $\texttt{surprisal}(w_t)$ のみがベースライン回帰モデルから覗かれる．単語頻度 $\texttt{freq}(w_t)$ は，Wiki-103 データ [24] に基づいている．単語長 $\texttt{length}(w_t)$ は文字数で求めている．

B　言語モデル

実験に用いたモデルを表 3 に示す．計算リソースの都合上，特定のモデルでは推論時に量子化を適用した．

C　データ

既存研究 [6, 25, 13] に従い，被験者内での平均読み時間を用いた．読み時間がゼロか，コーパス内で 3 標準偏差を超えるデータポイントは除外した．また，文頭と文末の単語も対象から除外した．

D　指示

実験で用いた指示を表 4 に示す．なお LLaMa-2 では，`Answer:` によって話者交代を明示的に示さない場合，指示の続きを補完する傾向にあったため，指示を多少変更した．

E　指示条件付き言語モデル

指示の変更により文補完の傾向が意図されたとおりに変わることを確認する．例として，異なる指示（1–3）の元で文を補完し，補完された文における係り受けの距離の分布を，図 E に示す．文法的に単純・複雑になるよう指示した場合に，たしかに統語距離が短く・長くなることが確認される．

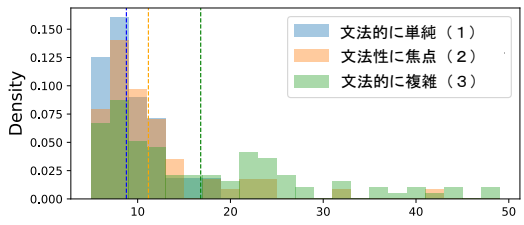

図 3　指示ごとの統語依存距離の分布 (> 5)

F　メタ言語的指示

表 5 と表 6 に，実際に用いたメタ言語的指示を示す．

表 4　実験で使用した指示

GPT3.5, Falcon の場合	LLaMA-2 の場合
Please complete the following sentence to make it as grammatically simple as possible: \n $w_0, \cdots, w_{t-1}$	Please generate a grammatically simple sentence as much as possible.\n\nAnswer: \n $w_0, \cdots, w_{t-1}$
Please complete the following sentence with a careful focus on grammar \n $w_0, \cdots, w_{t-1}$	Please generate a sentence with a careful focus on grammar.\n\nAnswer: \n $w_0, \cdots, w_{t-1}$
Please complete the following sentence to make it as grammatically complex as possible: \n $w_0, \cdots, w_{t-1}$	Please generate a grammatically complex sentence as much as possible.\n\nAnswer: \n $w_0, \cdots, w_{t-1}$
Please complete the following sentence using the simplest vocabulary possible: \n $w_0, \cdots, w_{t-1}$	Please generate a sentence using the simplest vocabulary possible.\n\nAnswer: \n $w_0, \cdots, w_{t-1}$
Please complete the following sentence with a careful focus on word choice \n $w_0, \cdots, w_{t-1}$	Please generate a sentence with a careful focus on word choice.\n\nAnswer: \n $w_0, \cdots, w_{t-1}$
Please complete the following sentence using the most difficult vocabulary possible: \n $w_0, \cdots, w_{t-1}$	Please generate a sentence using the most difficult vocabulary possible.\n\nAnswer: \n $w_0, \cdots, w_{t-1}$
Please complete the following sentence in a human-human-like manner. It has been reported that human ability to predict next words is weaker than language models and that humans often make noisy predictions, such as careless grammatical errors. \n $w_0, \cdots, w_{t-1}$	Please generate a sentence in a human-like manner. It has been reported that human ability to predict next words is weaker than language models and that humans often make noisy predictions, such as careless grammatical errors.\n\nAnswer: \n $w_0, \cdots, w_{t-1}$
Please complete the following sentence. We are trying to reproduce human reading times with the word prediction probabilities you calculate, so please predict the next word like a human. It has been reported that human ability to predict next words is weaker than language models and that humans often make noisy predictions, such as careless grammatical errors. \n $w_0, \cdots, w_{t-1}$	Please generate a sentence. We are trying to reproduce human reading times with the word prediction probabilities you calculate, so please predict the next word like a human. It has been reported that human ability to predict next words is weaker than language models and that humans often make noisy predictions, such as careless grammatical errors.\n\nAnswer: \n $w_0, \cdots, w_{t-1}$
Please complete the following sentence: \n $w_0, \cdots, w_{t-1}$	Please generate a sentence.\n\nAnswer: \n $w_0, \cdots, w_{t-1}$

表 5　メタ言語的指示 1（処理負荷を質問）

Suppose humans read the following sentence: "'No, it's fine. I love it,' said Lucy knowing that affording the phone had been no small thing for her mother."
List the tokens and their IDs in order of their reading cost (high to low) during sentence processing.
Token ID:
0: 'No,, 1: it's, 2: fine., 3: I, 4: love, 5: it,', 6: said, 7: Lucy, 8: knowing, 9: that, 10: affording, 11: the, 12: phone, 13: had, 14: been, 15: no, 16: small, 17: thing, 18: for, 19: her, 20: mother.,
Answer:
20: mother., 10: affording, 6: said, 11: the, 0: 'No,, 7: Lucy, 1: it's, 9: that, 17: thing, 5: it,', 2: fine., 15: no, 14: been, 3: I, 13: had, 8: knowing, 12: phone, 19: her, 16: small, 4: love, 18: for,

事例 2（省略）
事例 3（省略）

Suppose humans read the following sentence: （対象とする文）
List the tokens and their IDs in order of their reading cost (high to low) during sentence processing.
Token ID:
（対象とする文における単語）
Answer:

表 6　メタ言語的指示 2（確率を質問）

Suppose you read the following sentence: "'No, it's fine. I love it,' said Lucy knowing that affording the phone had been no small thing for her mother."
List the tokens and their IDs in order of their probability in context (low to high).
Token ID:
0: 'No,, 1: it's, 2: fine., 3: I, 4: love, 5: it,', 6: said, 7: Lucy, 8: knowing, 9: that, 10: affording, 11: the, 12: phone, 13: had, 14: been, 15: no, 16: small, 17: thing, 18: for, 19: her, 20: mother.,
Answer:
0: 'No,, 10: affording, 8: knowing, 12: phone, 4: love, 5: it,', 7: Lucy, 15: no, 13: had, 17: thing, 1: it's, 6: said, 2: fine., 20: mother., 11: the, 18: for, 16: small, 9: that, 19: her, 3: I, 14: been,

exemplar 2

exemplar 3

Suppose you read the following sentence: ¡TARGET SENT¿
List the tokens and their IDs in order of their probability in context (low to high).
Token ID:
¡TOKENS FROM TARGET SENT¿
Answer: