Psychometric Predictive Power of Large Language Models

Tatsuki Kuribayashi[1] **Yohei Oseki**[2] **Timothy Baldwin**[1,3]
[1]MBZUAI [2]The University of Tokyo [3]The University of Melbourne
{tatsuki.kuribayashi,timothy.baldwin}@mbzuai.ac.ae
oseki@g.ecc.u-tokyo.ac.jp

Abstract

Instruction tuning aligns the response of large language models (LLMs) with human preferences. Despite such efforts in human–LLM alignment, we find that instruction tuning does not always make LLMs human-like from a cognitive modeling perspective. More specifically, next-word probabilities estimated by instruction-tuned LLMs are often worse at simulating human reading behavior than those estimated by base LLMs. In addition, we explore prompting methodologies for simulating human reading behavior with LLMs. Our results show that prompts reflecting a particular linguistic hypothesis improve psychometric predictive power, but are still inferior to small base models. These findings highlight that recent advancements in LLMs, i.e., instruction tuning and prompting, do not offer better estimates than direct probability measurements from base LLMs in cognitive modeling. In other words, pure next-word probability remains a strong predictor for human reading behavior, even in the age of LLMs.

https://github.com/kuribayashi4/llm-cognitive-modeling

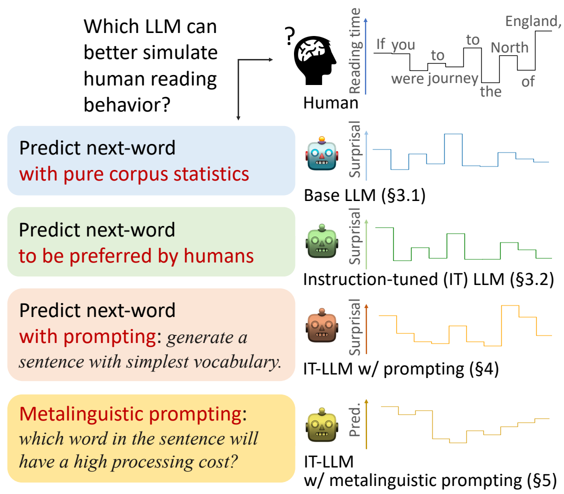

Figure 1: Comparing the "reading behavior" of humans and LLMs, i.e., reading time from humans is compared with surprisal from LLMs (§2.1). We investigate which surprisal values estimated by: (i) base LLMs, (ii) instruction-tuned (IT) LLMs, (iii) IT-LLMs with prompting, or (iv) IT-LLMs with metalinguistic prompting can better simulate human reading time.

1 Introduction

Aligning computational models with human perception/cognition has historically been a pivotal approach to understanding humans (Shapiro, 2003). With this in mind, computational psycholinguistics has investigated the model of human sentence processing (Crocker, 2007) and recently found an intriguing correlation between next-word probabilities from language models (LMs) and human reading behavior—the less predictable a word is, the greater the cognitive load (e.g., longer reading time) humans exhibit—suggesting the expectation-based account of human sentence processing (Levy, 2008; Smith and Levy, 2013). Based on this finding, the field has further investigated which types of models/algorithms can compute probabilities better aligned with human reading behavior (Figure 1; Hale (2001); Goodkind and Bicknell (2018); Wilcox et al. (2020); Oh et al. (2021); Kuribayashi et al. (2022); *inter alia*).

In the field of natural language processing (NLP), in contrast, large language models (LLMs) tuned to human-preferred responses (e.g., GPT-3.5) improve in performance across a wide range of applications (Ouyang et al., 2022). Given the increasing prevalence of such human-aligned, *instruction-tuned* LLMs (IT-LLMs), the following computational psycholinguistic question naturally arises: *do IT-LLMs successfully simulate human reading behavior in terms of predicted surprisal?* The answer to this question is not immediately obvious. On the one hand, the answer might be *yes* since these are tuned to human-preferred responses (Zhang et al., 2023), which will be, broadly

Findings of the Association for Computational Linguistics: NAACL 2024, pages 1983–2005

speaking, more aligned with human-like expectations of upcoming information during reading, e.g., not expect fabricated/hallucinated information during reading (Grice, 1975; Askell et al., 2021). Moreover, some IT-LLMs employ reinforcement learning from human feedback (RLHF); such a scenario of language learning through (approximate) human feedback is more plausible than through text alone (Alishahi, 2010), and thus may enhance their cognitive plausibility.

On the other hand, IT-LLMs are a step beyond base LMs—pure statistical models of plausible text based on large natural language corpora—in the sense that IT-LLMs are tuned to specific human-preferred responses and may suffer from amplified reporting biases. If the core of human sentence processing is explained by next-word probabilities (Levy, 2008; Smith and Levy, 2013), instruction tuning will be unnecessary in simulating human reading behavior. Based on this logic, the answer would be *no*.

Our experiments show that IT-LLMs frequently yield worse psychometric predictive power (PPP) for human reading behavior than base LLMs (§3). This suggests that the current paradigm of instruction tuning is irrelevant to modeling human sentence processing; IT-LLMs are not specially aligned with human language processing, at least based on reading time modeling, despite their ultimate goal of human–machine alignment.

Furthermore, we address open questions regarding prompting in cognitive modeling: (i) Can prompting result in probabilities from IT-LLMs being more aligned with human reading behavior? and (ii) Which is better at simulating human reading behavior, using direct probability measurements or directly asking IT-LLMs about the processing cost, e.g., *metalinguistic prompting* (Hu and Levy, 2023)? For the first question, we find prompts in line with so-called "good-enough" human sentence processing (Ferreira and Lowder, 2016) to work well. However, these are still worse than smaller base LLMs (§4). For the second question, we find metalinguistic prompting to be inferior to direct probability measurement in terms of PPP (§5).

In sum, despite the recent advancements in IT-LLMs and prompting, they do not currently offer better estimates of human reading behavior than simple probability measurements from base LLMs. This also underlines the value of access to probabilistic outputs for closed-source LLMs to further the study of cognitive modeling.

2 Simulating human reading behavior

2.1 Linking hypothesis

It has been reported that the word-by-word processing cost for humans, typically measured by reading time (RT), can be explained by the surprisal of a word $h_{t,\theta}(w)$ in context $\boldsymbol{w}_{<t} = [w_0, \cdots, w_{t-1}]$, computed by a model θ (Hale, 2001; Levy, 2008; Smith and Levy, 2013):

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

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

To gauge the advantage of the surprisal factor in reading time modeling, we train two nested regression models[1] (Eq. 1) with and without the surprisal factor in addition to the $\mathrm{baselines}(w_t)$ factors.[2] Then, we report the psychometric predictive power (PPP), which is defined as the increase in the per-token average of the log-likelihood of the regression model due to the added surprisal factor. A high PPP indicates the effectiveness of the surprisal factor in simulating human reading behavior. Our interest in this paper is to find the model θ that leads to a higher PPP.

Following existing studies (Roark et al., 2009; van Schijndel and Linzen, 2019; Pimentel et al., 2022), we also examine other variants of Eq. 1 by replacing the surprisal factor $h_{t,\theta}(w)$ with the expected value of surprisal $\mathrm{H}_\theta(W_t)$, in the form of: (i) contextualized Shannon entropy (Shannon, 1948); and (ii) its generalization called contextualized Rényi entropy $\mathrm{H}_{\alpha,\theta}(W_t)$ (Rényi, 1961):

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

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

Here, vocabulary set W_t is approximated by the

[1] We used `statsmodels` (Seabold and Perktold, 2010).

[2] We used the following formulation: $\texttt{RT}(w_t) \sim \texttt{surprisal}(w_t) + \texttt{surprisal}(w_{t-1}) + \texttt{surprisal}(w_{t-2}) + \texttt{length}(w_t) + \texttt{freq}(w_t) + \texttt{length}(w_{t-1}) + \texttt{freq}(w_{t-1}) + \texttt{length}(w_{t-2}) + \texttt{freq}(w_{t-2})$. The $\texttt{surprisal}(w_t)$ factor is excluded in the baseline regression model. The $\texttt{freq}(w_t)$ is quantified based on Wiki-103 data (Stephen et al., 2017) with logarithmic conversion. The $\texttt{length}(w_t)$ is the character length of the word. Adding an interaction term $\texttt{length}(w_i)\texttt{*}\texttt{freq}(w_i)$ did not alter our findings, and thus we use the simpler independent model.

model's subword vocabulary set.[3] For the Rényi entropy, we set $\alpha = 0.5$ based on the results of Pimentel et al. (2022); Liu et al. (2023), noting that Shannon entropy is a special case of Rényi entropy with $\alpha = 1$. Note that we limit $\boldsymbol{w}_{<t}$ to intra-sentential context since we are interested in sentence-level language processing in this study. Word boundaries are based on the reading-time corpus; if a word consists of several subwords, cumulative surprisal is computed.

2.2 Experimental settings

Models: We examined 26 LLMs as candidate models θ to compute information-theoretic values: six LLaMA-2 (Touvron et al., 2023), four Falcon (Almazrouei et al., 2023), four GPT-2 (Radford et al., 2019), four GPT-3/3.5 (Ouyang et al., 2022)[4], and eight OPT (Zhang et al., 2022) models with different sizes and instruction tuning settings (see Appendix A for details). Among them, GPT-3.5, two LLaMA-2, and two Falcon models are IT-LLMs (models with ✓ in the "IT" column in Table 1), and the others are "base LLMs." More specifically, GPT-3.5 D2 is trained through supervised fine-tuning, Falcon IT-LLMs are also trained via a particular supervised-tuning approach (Xu et al., 2023)[5], and GPT-3.5 D3 and LLaMA-2 IT-LLMs employ RLHF. Note that entropy metrics are omitted from the GPT-3/3.5 results since their APIs do not provide the probability distribution over the entire vocabulary.

Data: We use two corpora: Dundee Corpus (DC: Kennedy et al. (2003)) and Natural Stories Corpus (NS: Futrell et al. (2018)). DC is eye-tracking data, where we use the first-pass duration as reading time, while NS is self-paced reading time data. Following recent studies (Wilcox et al., 2020, 2021; Pimentel et al., 2022), we averaged the reading times for each word across different human subjects. We excluded data points with a reading time of zero or beyond three standard deviations. We also excluded the sentence-initial/final words since IT-LLMs tend to predict special phrases (e.g., *Sure, here is my answer...*) at the sentence-initial position, and sentence-final words potentially have a confounding influence (Rayner et al., 2000; Meister et al., 2022).

		DC				NS			
Model	IT	$h\uparrow$	H $\uparrow$	$\mathrm{H}_{0.5}\uparrow$	PPL $\downarrow$	$h\uparrow$	H $\uparrow$	$\mathrm{H}_{0.5}\uparrow$	PPL $\downarrow$
LMA-2 7B		10.33	8.58	13.45	76.40	6.41	3.06	9.97	45.21
LMA-2 7B	✓	8.97	5.57	12.03	153.46	7.07	2.42	8.33	63.74
LMA-2 13B		9.44	8.04	13.77	75.28	5.44	2.44	9.23	41.62
LMA-2 13B	✓	9.13	5.30	11.97	123.35	5.93	1.99	7.53	56.05
LMA-2 70B		8.21	5.14	10.47	78.28	4.51	1.80	6.79	37.61
LMA-2 70B	✓	8.67	4.53	10.67	112.07	5.60	1.75	7.34	52.05
Falcon 7B		9.08	7.75	11.81	97.86	7.61	3.95	12.17	49.64
Falcon 7B	✓	11.18	8.57	12.31	131.53	8.54	4.38	12.63	62.99
Falcon 40B		8.53	6.93	10.99	77.72	5.35	2.41	9.36	41.46
Falcon 40B	✓	9.06	6.76	10.43	92.53	5.49	2.89	8.49	47.27
GPT-3 B2		12.47	-	-	108.77	10.58	-	-	57.91
GPT-3 D2		9.93	-	-	79.65	6.45	-	-	44.79
GPT-3.5 D2	✓	9.35	-	-	72.95	5.30	-	-	38.23
GPT-3.5 D3	✓	8.91	-	-	84.17	5.83	-	-	44.38
GPT-2 177M		15.23	12.32	15.55	209.37	15.61	10.20	18.19	93.81
GPT-2 355M		9.63	11.20	15.37	222.17	13.62	8.91	16.96	75.67
GPT-2 774M		10.98	9.66	14.79	165.81	12.04	7.01	14.52	66.87
GPT-2 1.5B		10.18	-	14.15	158.75	10.94	6.99	14.69	65.14
OPT 125M		15.65	13.72	17.18	231.80	15.54	12.27	19.41	109.11
OPT 350M		14.81	11.89	16.07	196.02	14.86	10.35	18.11	94.51
OPT 1.3B		10.51	10.16	15.55	160.95	11.81	7.43	16.53	67.59
OPT 2.7B		9.52	9.65	14.38	150.78	11.66	6.60	15.51	63.98
OPT 6.7B		9.43	9.06	13.63	130.01	9.59	5.56	13.64	57.86
OPT 13B		9.06	8.57	13.15	130.44	9.51	4.96	12.84	56.74
OPT 30B		9.62	8.58	13.17	119.42	8.55	4.16	10.39	54.91
OPT 66B		10.30	7.42	12.73	94.15	7.78	4.33	11.92	49.11

Table 1: The PPL and PPP scores of tested LMs. The "IT" column denotes whether instruction tuning is applied. The columns h, H, and $\mathrm{H}_{0.5}$ indicate surprisal, Shannon entropy, and Rényi entropy ($\alpha = 0.5$), respectively. The colors of cells for IT-LLMs indicate if the PPP increased or decreased relative to its base version (GPT-3.5 models are compared to GPT-3s)."LMA-2" denotes the LLaMA-2 family.

3 Experiment 1: PPP of LLMs

We first observe the PPP of base LLMs (§3.1) and then analyze the PPP of IT-LLMs (§3.2). We explore prompting in §4 and §5.

3.1 Reproducing previous findings

Table 1 shows the PPP and perplexity (PPL)[6] of each LLM. We first examine whether we are able to reproduce results from existing studies.

Surprisal theory: Across all the settings of {model×corpus×metric}, information-theoretic

[3] A set of entire words in natural language can inherently be infinite. See Appendix B in Pimentel et al. (2022) for the details about subword-based entropy.

[4] GPT-3 B2/D2 denotes `bebbage-002` and `davinci-002`, respectively. GPT-3.5 D2/D3 denotes `text-davinci-002` and `text-davinci-003`, respectively.

[5] Further details are not clear at the time of writing this paper. We solely rely on the description here: `https://huggingface.co/tiiuae/falcon-40b-instruct`

[6] We measure the PPL of an LLM as average surprisal power of 2: $2^{\frac{1}{N}\sum_t h_{t,\theta}(w)}$ over the reading-time-annotated units; this ensures comparable PPL scores across LLMs with different tokenizers.

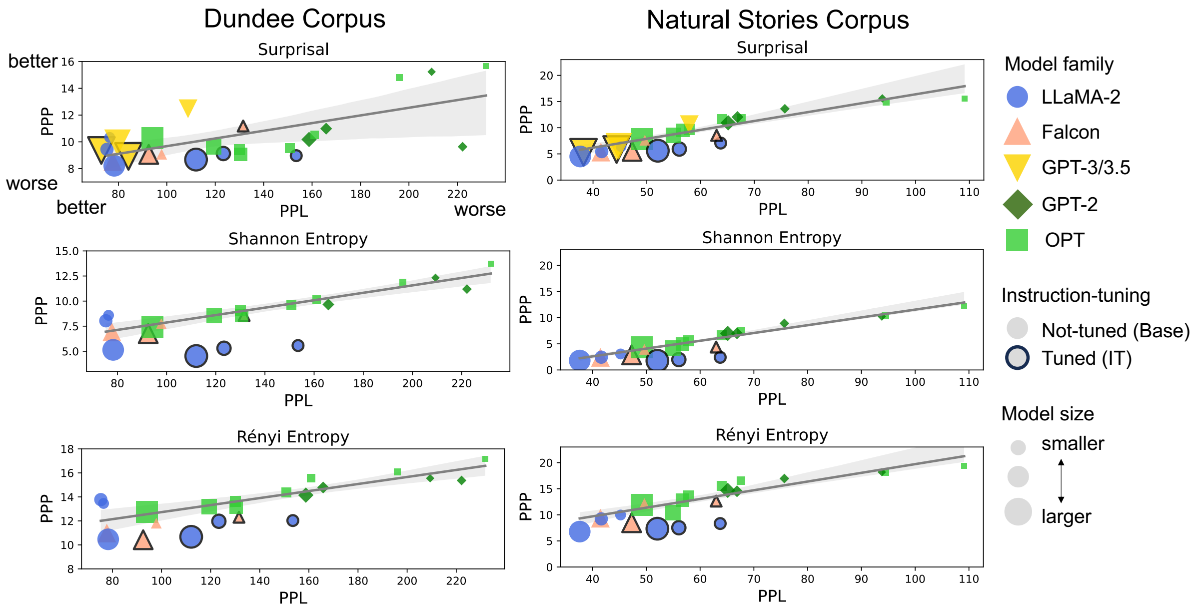

Figure 2: The relationship between PPL and PPP (see exact scores in Table 1). Each point corresponds to each LLM, and those with a black edge line are IT-LLMs. The regression line is estimated by base LLMs, and the colored area presents a 95% confidence interval. IT-LLMs were relatively poor (below the line) at balancing PPL and PPP.

metrics are effective in simulating reading time as shown in existing studies (Shain et al., 2022, *inter alia*). The PPP scores were significantly positive (F-test, $p < 0.001$), and the coefficients of interest were also significantly larger than zero (one-sample two-sided t-test, $p < 0.001$).

Advantage of Rényi entropy: Rényi entropy with $\alpha = 0.5$ ($\mathrm{H}_{0.5}$) is consistently better at simulating human reading behavior (Table 1) than the other metrics (h and H) in all settings. Such an advantage has been reported with GPT-2 models (Pimentel et al., 2022; Liu et al., 2023); we further show that this generalizes to other LLMs.

PPP–PPL inverse scaling: While early work showed that better PPL leads to better PPP (Roark et al., 2009; Frank and Bod, 2011; Goodkind and Bicknell, 2018), more recent work using neural LLMs has reported the opposite, inverse relationship between PPP and PPL when using base LLMs—the **worse** the PPL is, the better the PPP is (Kuribayashi et al., 2022; Shain et al., 2022; de Varda and Marelli, 2023; Oh and Schuler, 2023), implying a "superhuman" ability of LLMs in next-word prediction. The results of Table 1 are mapped onto the PPP and PPL axes in Figure 2. Each point corresponds to an LLM, and those with black-edged outlines correspond to IT-LLMs (their results are analyzed in §3.2). The regression lines of the PPP–PPL relationship are estimated only by base LLMs; these lines replicate the inverse scaling effect, i.e., better PPL leads to worse PPP. Pearson's correlation between PPP and PPL was positive in all settings (r was within 0.55–0.95 with $p < 0.05$). Our results confirm that such a relationship holds even when using LLaMA-2, Falcon, and GPT-3/3.5 models and entropy metrics, extending the results of earlier studies (Oh and Schuler, 2023).

3.2 The effects of instruction tuning

Instruction tuning often hurts PPP: The green and red cells in Table 1 indicate the positive and negative effects of instruction tuning, respectively, based on comparison of the base model and its instruction-tuned variant (e.g., LLaMA-2 7B vs. LLaMA-2 7B instruct). There are no consistent positive or negative effects and thus no evidence that instruction tuning causes LLMs to be more cognitively plausible in terms of cognitive modeling. More specifically, LLaMA-2 and GPT-3.5 models tend to degrade under instruction tuning, while Falcon IT-LLMs show a somewhat positive effect. Falcon family IT-LLMs employ a supervised-tuning approach (Xu et al., 2023) rather than RLHF, suggesting that RLHF might lead to drastic negative effects. Nevertheless, there could be several confounding factors, e.g., the base model architecture, training regimen, and instruction-tuning data/policies, motivating future work to investigate this effect in a more controlled manner.

Worse PPP compared to base LLMs with equivalent PPL: We additionally report a consistent tendency that IT-LLMs yielded poorer PPP than the base LLMs with equivalent PPL, on top of the reported PPP-PPL trade-off (Oh and Schuler, 2023). Figure 2 shows that IT-LLMs (points with black-edged outlines) are positioned below the PPP–PPL regression line estimated by the base LLMs across all metrics and corpora. Specifically, IT-LLMs yield poor PPL scores, presumably due to their objective no longer being pure language modeling. Worse still, they also yield worse PPP scores than the expected good values based on the estimated inverse PPP-PPL relationship (gray lines in Figure 2). Specifically, 32 results out of 34 {IT-LLM×metric×corpus} settings are below the regression line. This is significantly more frequent than chance $\pi = 0.5$ based on a two-sided binomial test ($p < 1\mathrm{e}-7$). That is, IT-LLMs struggle to balance PPP and PPL.[7]

Discussion: These results indicate that instruction tuning **does not enhance the simulation of human reading behavior**, despite it being intended to make LLMs more human-aligned. There are at least two hypotheses for why this should be the case: (i) instruction tuning with manipulated text amplifies reporting biases in training data and corrupts the LLM's language statistics built on naturally occurring text during pretraining; and (ii) the instruction-tuning objective is misaligned with human reading, e.g., IT-LLMs are trained to correctly predict responses over a huge range of questions, well beyond the capacity of a single human subject. In other words, our results support the surprisal theory that links human sentence processing efforts with pure word surprisal (Levy, 2008; Smith and Levy, 2013; Shain et al., 2022).

4 Experiment 2: prompting

The use of IT-LLMs stimulates an additional question—*does prompt tuning lead to better PPP for IT-LLMs?* One can control the *prior belief* of LLMs about upcoming words through the prompt, e.g., *complete the sentence to make it grammatically simple.* Analyzing effective prompt types may shed light on underlying bias in human expectation-based reading, i.e., which kinds of words are more expected by humans.

ID	Prompt
Syn↓	Please complete the following sentence to make it as grammatically simple as possible: \n $w_0, \cdots, w_{t-1}$
Syn→	Please complete the following sentence with a careful focus on grammar: \n $w_0, \cdots, w_{t-1}$
Syn↑	Please complete the following sentence to make it as grammatically complex as possible: \n $w_0, \cdots, w_{t-1}$
Lex↓	Please complete the following sentence using the simplest vocabulary possible: \n $w_0, \cdots, w_{t-1}$
Lex→	Please complete the following sentence with a careful focus on word choice: \n $w_0, \cdots, w_{t-1}$
Lex↑	Please complete the following sentence using the most difficult vocabulary possible: \n $w_0, \cdots, w_{t-1}$
Task1	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 $w_0, \cdots, w_{t-1}$
Task2	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 $w_0, \cdots, w_{t-1}$
Base	Please complete the following sentence: \n $w_0, \cdots, w_{t-1}$

Table 2: Our examined prompts. The IDs shown in the first column are also used in Tables 3 and 4.

Settings: We examine the nine prompts shown in Table 2 to linguistically bias the LLM-computed information-theoretic values.[8] The first six prompts focus on syntactic and lexical complexity, based on longstanding interest in syntactic and lexical biases in expectation-based reading (Roark et al., 2009; Frank and Bod, 2011). The "Task1" and "Task2" prompts inform IT-LLMs of the task-specific objective of our experiments. A prompt is appended immediately before the context words ($\boldsymbol{w}_{<t}$ in Eq. 2) when computing the information-theoretic values. That is, we now use **prompt-conditioned surprisal, Shannon entropy, and Rényi entropy** ($\alpha = 0.5$) values with a given prompt r:

[7]To handle the concern of IT-LLMs simply being confused for a given sentence fragment $\boldsymbol{w}_{<t}$ in isolation, we re-conducted the experiments with the explicit instruction to predict the next word, yielding results consistent with the original; that is, the results hold up even after instructing IT-LLMs to behave as base LLMs (Appendix B).

[8]We used a slightly different prompting format for LLaMA-2s (see Appendix C.1).

	dependency length ↑					sentence length ↑				
	LLaMA-2			Falcon		LLaMA-2			Falcon	
Prompt	7B	13B	70B	7B	40B	7B	13B	70B	7B	40B
Syn↓	2.10	2.43	2.51	2.27	2.48	12.3	14.6	16.4	12.8	15.7
Syn→	2.90	2.87	2.68	2.65	3.47	24.9	22.9	21.2	18.4	18.1
Syn↑	**3.45**	**3.29**	**3.31**	2.86	**3.49**	**43.9**	**44.0**	**45.5**	**23.8**	**33.5**
Lex↓	2.31	2.40	2.46	2.28	3.01	13.4	13.1	14.7	13.5	15.7
Lex→	2.95	3.29	3.05	2.58	3.33	32.4	28.0	25.2	17.5	19.9
Lex↑	3.08	3.06	3.24	2.86	3.23	35.2	36.5	33.8	24.6	27.6
Task1	2.87	2.94	2.82	2.70	2.87	27.1	28.2	28.6	21.1	26.2
Task2	2.99	2.80	3.07	2.79	2.82	23.6	20.4	18.3	21.2	21.4
Base	2.65	2.47	2.77	**2.91**	3.00	17.9	17.4	19.3	20.4	18.9

(a) Syntactic complexity

	log word frequency ↓					word length ↑				
	LLaMA-2			Falcon		LLaMA-2			Falcon	
Prompt	7B	13B	70B	7B	40B	7B	13B	70B	7B	40B
Syn↓	4.77	4.89	4.69	4.87	4.87	3.97	4.08	4.23	3.78	4.13
Syn→	4.55	4.67	4.63	4.79	4.84	4.44	4.53	4.46	3.94	4.39
Syn↑	4.43	4.44	4.55	4.64	4.70	4.60	4.68	4.67	4.84	4.35
Lex↓	4.73	4.80	4.73	4.79	4.85	4.16	3.62	4.22	4.03	4.11
Lex→	4.55	4.76	4.57	4.50	4.84	4.49	4.51	4.54	4.59	4.31
Lex↑	**3.90**	**4.09**	**3.80**	**4.37**	**4.44**	**5.16**	**4.98**	**5.27**	**4.99**	**4.95**
Task1	4.75	4.73	4.68	4.65	4.78	4.61	4.31	4.47	4.94	4.27
Task2	4.80	4.75	4.83	4.87	4.80	4.28	4.14	4.23	4.37	4.46
Base	4.77	4.86	4.87	4.93	4.91	4.37	4.21	4.23	4.15	4.16

(b) Lexical complexity. Stopwords are excluded when calculating the average of logarithmic word frequencies.

Table 3: Statistics of sentences generated with different prompts and IT-LLMs. The highest values of dependency, sentence, and word length, and the lowest value of log word frequency for each model are highlighted.

$$h_{t,\theta}(w,r) := -\log_2 p_\theta(w|\boldsymbol{w}_{<t},r) \quad , \tag{5}$$

$$\mathrm{H}_\theta(W_t,r) := \mathop{\mathbb{E}}_{w\sim p(\cdot|\boldsymbol{w}_{<t})} h_{t,\theta}(w,r) \quad , \tag{6}$$

$$\mathrm{H}_{\alpha,\theta}(W_t,r) := \lim_{\gamma\to\alpha}\frac{1}{1-\gamma}\log_2\sum_{w\in W} p_\theta(w|\boldsymbol{w}_{<t},r)^\gamma \quad . \tag{7}$$

4.1 Preliminary: does prompting control next-word prediction?

To ensure that prompting does indeed induce the intended changes in the next-word distribution, we first analyze sentences generated with different prompts using LLaMA-2 and Falcon IT-LLMs. To diversify the input context, we first extract 20 sentences from the DC dataset and append their first five words to each of the nine prompts, resulting in 180 (=20×9) input contexts. Then, we generate 180 sentences by feeding the respective input contexts to the IT-LLMs. To measure the syntactic and lexical biases in text generation, we report the averaged dependency length, sentence length, logarithmic word frequency, and word length of the sentences generated with each prompt. Appendix C.2 provides more details of this preliminary analysis.

Prompting controls next-word prediction as intended: Table 3 shows the statistics of sentences generated by different prompts. As intended, syntactically complex, long sentences with long syntactic dependencies are generated when the model is instructed to make sentences grammatically complex (Syn↑; Table 3a), and vice versa (Syn↓; Table 3a). The lexical biases are also appropriately injected by prompting: the Lex↑ (Lex↓) prompts result in sentences with less (more) frequent and longer (shorter) words (Table 3b). On the basis of this, we can infer that LLMs have some ability to bias sentence completion prediction based on linguistic instructions. This finding itself opens up the potential of prompting as a way of controlling the linguistic bias in information-theoretic values rather than, e.g., training differently-biased LLMs from scratch as typically done in computational psycholinguistics research (Frank and Bod, 2011). Note that, nevertheless, LLMs are not able to perfectly separate the two orthogonal dimensions of syntactic and lexical complexity in text generation with prompts. For example, Lex↓ and Lex↑ prompts somewhat affect the syntactic complexity of generated sentences (Table 3a) as well. Progress in controlled text generation should mitigate this effect (Zhou et al., 2023).

4.2 Results: PPP of prompt-conditioned LLMs

Effective prompts: Table 4 shows the averaged PPP against each prompt in each IT-LLM family: instruction-tuned LLaMA-2 (7B, 13B, 70B), Falcon (7B, 40B), and GPT-3.5 (D2 and D3) models. We also report two baseline results obtained with the "Base" prompt (Table 2) and without prompting. We observe the following: (i) some prompts, typically, Syn↓ and Task2, lead LLMs to achieve better PPP than baselines; (ii) such effective prompting strategies are generally consistent across corpora; and (iii) instructions to use simple grammar/vocabulary work better than those to use complex grammar/vocabulary. In terms of the first finding, some prompt-conditioned LLMs exhibit better PPP than the baselines, especially based on the entropy measurements (H and $\mathrm{H}_{0.5}$). For the second finding, particular prompts, e.g., Syn↓ and

	DC							NS						
	$h\uparrow$			H $\uparrow$		$\mathrm{H}_{0.5}\uparrow$		$h\uparrow$			H $\uparrow$		$\mathrm{H}_{0.5}\uparrow$	
Prompt	LMA-2	Falcon	GPT-3.5	LMA-2	Falcon	LMA-2	Falcon	LMA-2	Falcon	GPT-3.5	LMA-2	Falcon	LMA-2	Falcon
Syn↓	8.20	**9.73**	6.20	5.94	**9.77**	12.30	**12.69**	3.84	7.38	**5.83**	2.57	4.60	8.78	13.30
Syn→	8.63	9.46	6.56	5.56	9.52	11.94	12.40	4.34	**7.46**	5.23	1.88	4.63	7.14	13.01
Syn↑	8.18	9.46	5.90	5.64	9.60	11.66	12.56	4.64	7.10	4.07	1.31	4.09	6.38	12.35
Lex↓	7.93	9.49	5.92	6.33	9.74	12.39	12.63	3.61	6.77	4.79	2.00	4.74	8.34	13.26
Lex→	8.10	9.16	**6.63**	5.18	9.18	11.34	12.28	3.79	7.30	5.30	1.37	4.21	6.36	12.75
Lex↑	7.79	8.81	5.77	5.34	9.29	11.44	12.34	3.66	6.01	3.88	1.12	3.65	6.00	11.53
Task1	8.83	8.82	5.54	5.99	8.86	12.39	12.35	4.29	5.46	2.95	2.61	4.73	**9.64**	13.42
Task2	**8.90**	9.16	5.95	**6.53**	9.13	**13.11**	12.42	**4.76**	5.99	3.23	**2.64**	**5.14**	9.43	**13.54**
Base	8.42	9.60	6.98	5.65	9.45	12.15	12.41	4.94	7.13	4.69	1.94	5.01	7.82	13.52
W/o	8.92	10.12	9.13	5.13	7.67	11.56	11.37	6.20	7.02	5.56	2.05	3.63	7.73	10.56

Table 4: The PPP scores when using each prompt (the highest scores other than baseline ones for each corpus/metric are in boldface). Scores are averaged in each IT-LLM family ("LMA-2" denotes LLaMA-2). The columns h, H, and $\mathrm{H}_{0.5}$ indicate surprisal, Shannon entropy, and Rényi entropy ($\alpha = 0.5$) settings, respectively. "W/o" denotes the setting without any prompt.

Task2, generally result in the best PPP on both the DC and NS datasets. Finally, for the third finding, prompts to use simple grammar/vocabulary (Syn↓ and Lex↓) lead to better PPP than the opposite instructions (Syn↑ and Lex↑). Note that, interestingly, the task-specific instructions also generally worked well, especially for the entropy measures. Detailed results of each model and prompt are given in Appendix C.3.

Discussion: Regarding the third finding of the advantage of prompts to use simple grammar/vocabulary, one plausible implication is that LLMs tend to predict more complex, wordy text completions than human expectations during reading, and this is alleviated somewhat by these prompts. Such an implied **simplicity bias** in effective prompts potentially lends support to the theory of so-called good-enough processing in human language processing (Ferreira and Lowder, 2016).

4.3 Analysis: PPP–PPL relationship

Prompt-conditioned LLMs underperform base LLMs with equivalent PPLs: Figure 3 shows the PPP and PPL of IT-LLMs conditioned on different prompts, including those listed in Appendix C.1, superimposed on the results from Figure 2. Similar to the results in §3, the prompt-conditioned LLMs (points with red-edged outlines) are under the PPP–PPL regression line estimated from base LLMs; 448 results out of 468 {prompt×model×metric×corpus} settings are below the line. This is significantly more frequent than chance $\pi = 0.5$ based on a two-sided binomial test ($p < 1\mathrm{e}-105$). This shows that the base LLMs set the empirical Pareto front with respect to the PPP–PPL trade-off against prompt-biased LLMs, and in other words, none of the prompt-conditioned IT-LLMs could outperform the small base LMs with the best PPP, such as GPT-2 small. That is, base LLMs are a strong baseline in reading time modeling.

5 Experiment 3: metalinguistic prompting

Instead of using direct probability measurements, one can also ask IT-LLMs about the processing cost of words via *metalinguistic* prompting (Hu and Levy, 2023), e.g., *Please estimate the cognitive load of this word in context*. We explore such an approach in this section.

Settings: Based on preliminary experiments, asking a model to predict ther eading time for each word via prompting does not work well. Thus, we make the problem simpler: rank words in a sentence in order of processing difficulty—*which word in this sentence incurs a higher cognitive load during reading? Please order the words (high to low cost).* We ask such a question to the model, then calculate Spearman's rank correlation between the estimated word processing costs and their actual reading time in each sentence.[9] Then, these correlation scores are averaged across the sentences in the reading time corpus; a high score indicates the model's estimated being more aligned with read-

[9]If the LLMs output a word not in the target sentence, or the output lacks particular words in the targeted sentence, these tokens are excluded in computing the correlation.

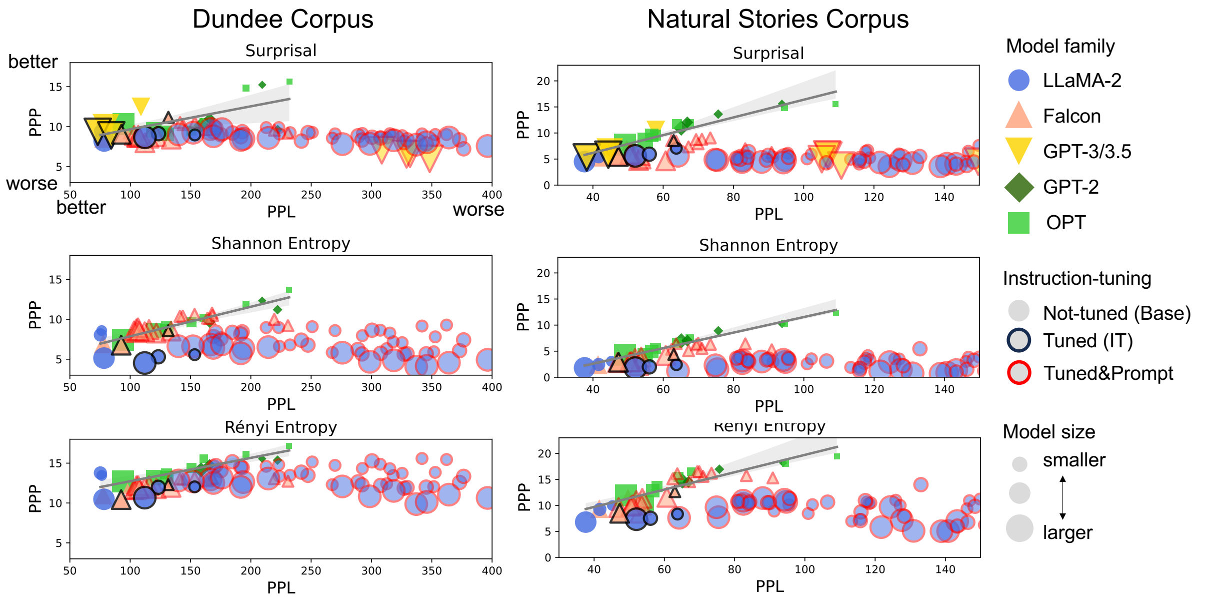

Figure 3: The PPL and PPP of LLMs with prompting are plotted at the top of Figure 2. Each point corresponds to a given combination of LLM and prompt, and those with red-edged outlines are IT-LLMs with a particular prompt. The PPP–PPL regression line is estimated by base LLMs, and the colored area presents the 95% confidence interval. IT-LLMs with prompting are poorer than base LLMs at balancing PPL and PPP.

ing time.[10] We employed a 3-shot setting,[11] and ran three different runs with different exemplars for prompting. The average and standard deviation scores across the three runs are reported (Table 5). We also calculated the rank correlation between surprisal and reading time as a baseline. LLaMA-2 70B is excluded due to computation resource limitations.

5.1 Results

Direct probability measurements outperform metalinguistic prompting: Table 5 shows the examined prompts and results (see Appendix D for the exact prompts). The standard surprisal-based method yields higher correlations than metalinguistic prompting methods, which yield near-zero correlations. Specifically, correlations from surprisal-based methods were significantly larger than those from metalinguistic prompting (two-sided Mann-Whitney U test, $p < 1e-8$) in both corpora. We suspect that the model simply failed in *ordering many items*; thus, we calculated the rank correlation coefficient by only using the first five words listed by the model, but the correlations were again nearly zero. Such inferiority is consistent with the results in grammaticality judgment tasks (Hu and Levy, 2023).

Prompt/method	Model	DC ↑	NS ↑
Suppose humans read the following sentence: (SENT). List the tokens in order of their reading cost (high to low) during sentence processing.	LMA-2 7B	0.09±0.02	-0.04±0.06
	LMA-2 13B	0.06±0.02	-0.03±0.06
	Falcon 7B	0.12±0.01	0.01±0.09
	Falcon 40B	0.03±0.04	−0.03±0.11
	GPT3.5 D2	0.05±0.03	0.05±0.03
	GPT3.5 D3	0.08±0.03	0.03±0.02
Suppose you read the following sentence: (SENT). List the tokens in order of their probability in context (low to high).	LMA-2 7B	0.05±0.06	0.00±0.02
	LMA-2 13B	0.04±0.03	0.06±0.04
	Falcon 7B	0.08±0.05	0.05±0.02
	Falcon 40B	0.02±0.07	0.13±0.10
	GPT3.5 D2	0.03±0.00	0.02±0.00
	GPT3.5 D3	−0.01±0.02	0.06±0.03
Surprisal-based estimation	LMA-2 7B	0.28	0.19
	LMA-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

Table 5: Average and standard deviation of rank correlation ρ between estimated cognitive load and reading time of words across three different runs. "LMA-2" denotes the LLaMA-2 family.

5.2 Analysis: metacognition of own surprisal

Gap between computed probability and response to metalinguistic prompts: *Are the estimates of word probability based on metalinguistic*

[10] 100 and 50 sentences are used for the DC and NS experiments, respectively. These are the first five sentences in each document in the respective corpus (DC/NS have 20/10 documents). We partially observed that these limited-scale experiments can approximate the full-scale experiments.

[11] Three exemplars are from the opposing corpus: the NS sentences were used for the DC experiments, and vice versa.

Model	DC ↑	NS ↑
LLaMA-2 7B	0.12±0.13	0.15±0.08
LLaMA-2 13B	0.02±0.10	0.06±0.07
Falcon 7B	0.15±0.08	0.30±0.09
Falcon 40B	0.09±0.09	0.17±0.00
GPT3.5 D2	0.15±0.02	0.22±0.07
GPT3.5 D3	0.18±0.05	0.24±0.02

Table 6: Rank correlation ρ between the word probability (rank) estimated by the prompt and the actual surprisal values computed by the corresponding model.

prompting consistent with their actual surprisal? To answer this question, we measure Spearman's rank correlation between the probabilities of words estimated via metalinguistic prompting (the second prompt in Table 5) and their actual surprisal values. The correlations are around 0.1–0.2 (Table 6), suggesting that metalinguistic prompting about word probability is again not an accurate measure of actual surprisal.

6 Related work

Simulating human reading behavior: Expectation-based accounts of human reading have been actively explored based on the linking hypothesis between surprisal and human reading behavior (Levy, 2008; Smith and Levy, 2013). To gain insights into black-box human sentence processing, analysis has been done on which types of models/algorithms better simulate human reading behavior (Hale, 2001; Demberg and Keller, 2008; Frank and Bod, 2011; Goodkind and Bicknell, 2018; Aurnhammer and Frank, 2019; Wilcox et al., 2020; Merkx and Frank, 2021; Kuribayashi et al., 2021; Noji and Oseki, 2021; Oh et al., 2021; Michaelov et al., 2021). It has been reported that those with specific properties, e.g., syntactic operations (Hale et al., 2018; Yoshida et al., 2021), memory limitations (Kuribayashi et al., 2022; Timkey and Linzen, 2023), and/or appropriate input units (Oh et al., 2021; Nair and Resnik, 2023) yield better fit to human reading behavior. Building on this body of work, we show that the current generation of IT-LLMs offers a somewhat poor predictor in cognitive modeling.

Prompt-based analysis of linguistic knowledge in LLMs: Given the rise of the prompting paradigm, testing the linguistic knowledge of LLMs via prompting has gained recent attention (Li et al., 2022; Hu and Levy, 2023; Katzir, 2023; Beguš et al., 2023; Dentella et al., 2023; Blevins et al., 2023). Prior work has pointed out their inferior ability at linguistic judgments under metalinguistic prompting to directly estimate probabilities (Hu and Levy, 2023; Dentella et al., 2023), and specifically Hu and Levy (2023) dubbed this discrepancy the so-called competence–performance distinction (Chomsky, 1965) of LLMs. This problem is also related to the calibration of model outputs (Kadavath et al., 2022). We revealed such degradation of metalinguistic prompting methods in simulating human reading behavior (§5).

Instruction tuning: Starting from the multitask fine-tuning of LMs (Wei et al., 2021; Sanh et al., 2022), *instruction-tuning*—aligning a model with human users—has played a crucial role in developing LLMs (Ouyang et al., 2022; Glaese et al., 2022). The objective of instruction tuning is, for example, making models *helpful*, *honest*, and *harmless* (Askell et al., 2021) in addition to just following the instructions. Notably, researchers may have been aligning LLMs to not the exact model of humans but rather a superhuman chat agent with instruction tuning; our results might reflect the paradox—pursuing human preferences has resulted in creating something different from humans.

Concurrent with this study, others have investigated the effects of instruction-tuning on the cognitive plausibility and linguistic knowledge of LLMs (Aw et al., 2023; Kauf et al., 2024). In particular, Aw et al. (2023) suggest that instruction tuning leads to a divergence in brain alignment and behavioral alignment, consistent with our work. The connection with our work requires further exploration.

7 Conclusion

We investigated the PPP of instruction-tuned LLMs, given their popularity in NLP. We found that IT-LLMs yield worse PPP than base LLMs with equivalent perplexity, demonstrating the ineffectiveness of current instruction tuning and (metalinguistic) prompting in simulating human reading behavior. One important area of future work is to explore why the current instruction-tuning paradigm is ineffective for modeling human reading behavior, and we highlighted the direction as *aligning LLMs with the human cognition/perception*, which has historically been a scientific approach to understanding humans, a.k.a. cognitive modeling, as well as with practical chat agents.

Limitations

Revealing why instruction tuning leads to a degradation in PPP is an important open question (§3.2). In particular, ablating reinforcement learning from human feedback (RLHF) via controlled experiments would be an interesting research direction. Unfortunately, the exact resources used in the instruction tuning of GPTs, LLaMA-2, and Falcon are not available, making it difficult to ablate the instruction-tuning scenario. Instead, training and evaluating separate LMs with different instruction tuning scenarios using publicly-available resources for instruction tuning (Conover et al., 2023; Taori et al., 2023) will be needed to further investigate our observation.

The scale of our experiments was limited across at least three dimensions. First, we only targeted the English language, although, to some degree, some language-dependent observations related to reading time have been obtained by prior work (Vasishth et al., 2010; Frank et al., 2016; Kuribayashi et al., 2021; Siegelman et al., 2022; Wilcox et al., 2023b; Pouw et al., 2023). Second, we tested only three families of IT-LLMs. Specifically, at the time of finalizing this paper, the GPT-3.5 models were no longer accessible, and newer versions do not offer the option to output the probability of the generated text; this motivates a focus on open LLMs for this line of research. Third, the variety of examined prompts was somewhat limited. Scaling up experiments with respect to these points is an obvious area for future work. Note that the use of pre-trained LMs poses the possibility of data leakage of source texts. Wilcox et al. (2023a) suggest that such leakage may not be a primary factor in characterizing PPP; thus, we tentatively put this concern aside in this study.

Ethical considerations

We do **not** in any way claim that the superiority of base LLMs in cognitive modeling, which potentially have harmful biases which are mitigated by instruction tuning, entails that human language processing also has inherently harmful biases. All we have shown in practice is a general macro-trend that surprisal from base LLMs has better PPP than instruction-tuned models, and such a potential bias in human language processing should be carefully inspected in separate work, e.g., as per Lior and Stanovsky (2023). We used some writing assistance tools, e.g., ChatGPT and Grammarly, in the writing of this paper, just to fix language errors.

Acknowledgement

We would like to thank Yova Kementchedjhieva, Go Inoue, and Ted Briscoe for their insightful feedback on earlier drafts of this paper. This work was supported in part by JST PRESTO Grant Number JPMJPR21C2, Japan.

References

Afra Alishahi. 2010. *Computational modeling of human language acquisition*. Morgan & Claypool Publishers.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Merouane Debbah, Etienne Goffinet, Daniel Heslow, Julien Launay, Quentin Malartic, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. Falcon-40B: an open large language model with state-of-the-art performance. *(not published)*.

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

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. 2023. Instruction-tuning aligns LLMs to the human brain. arXiv cs.CL/2312.00575.

Gašper Beguš, Maksymilian Dąbkowski, and Ryan Rhodes. 2023. Large linguistic models: Analyzing theoretical linguistic abilities of LLMs. `https://ling.auf.net/lingbuzz/007269`. Accessed: 2023-10-25.

Terra Blevins, Hila Gonen, and Luke Zettlemoyer. 2023. Prompting language models for linguistic structure. In *Proceedings of ACL 2023*, pages 6649–6663.

Noam Chomsky. 1965. *Aspects of the Theory of Syntax*. MIT Press.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. 2023. Free Dolly: Introducing the world's first truly open instruction-tuned LLM.

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

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

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

Vittoria Dentella, Elliot Murphy, Gary Marcus, and Evelina Leivada. 2023. Testing AI performance on less frequent aspects of language reveals insensitivity to underlying meaning. *arXiv preprint arXiv:2302.12313*.

Fernanda Ferreira and Matthew W Lowder. 2016. Prediction, information structure, and Good-Enough language processing. In Brian H Ross, editor, *Psychology of Learning and Motivation*, volume 65, chapter 6, pages 217–247. Academic Press.

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, Thijs Trompenaars, and Shravan Vasishth. 2016. Cross-linguistic differences in processing double-embedded relative clauses: Working-memory constraints or language statistics? *Cogn. Sci.*, 40(3):554–578.

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

Amelia Glaese, Nat McAleese, Maja Trębacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, et al. 2022. Improving alignment of dialogue agents via targeted human judgements. *arXiv preprint arXiv:2209.14375*.

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.

Herbert P Grice. 1975. Logic and conversation. In *Speech acts*, pages 41–58. Brill.

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

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

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

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Roni Katzir. 2023. Why large language models are poor theories of human linguistic cognition. a reply to piantadosi (2023). `https://ling.auf.net/lingbuzz/007190`. Accessed: 2023-10-25.

Carina Kauf, Emmanuele Chersoni, Alessandro Lenci, Evelina Fedorenko, and Anna A Ivanova. 2024. Comparing plausibility estimates in base and Instruction-Tuned large language models. arXiv cs.CL/2403.14859.

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

Tatsuki Kuribayashi, Yohei Oseki, 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.

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

Jiaoda Li, Ryan Cotterell, and Mrinmaya Sachan. 2022. Probing via prompting. In *Proceedings of NAACL 2022*, pages 1144–1157.

Gili Lior and Gabriel Stanovsky. 2023. Comparing humans and models on a similar scale: Towards cognitive gender bias evaluation in coreference resolution. In *Proceedings of CogSci 2023*, volume 45.

Tong Liu, Iza Škrjanec, and Vera Demberg. 2023. Improving fit to human reading times via temperature-scaled surprisal. arXiv cs.CL/2311.09325.

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 A Michaelov, Megan D Bardolph, Seana Coulson, and Benjamin Bergen. 2021. Different kinds of cognitive plausibility: why are transformers better than RNNs at predicting N400 amplitude? In *Proceedings of CogSci 2021*, volume 43.

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

Hiroshi Noji and Yohei Oseki. 2021. Effective batching for recurrent neural network grammars. In *Findings of ACL-IJCNLP 2021*, pages 4340–4352, Online.

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

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

OpenAI. 2023. GPT-4 Technical Report. *arXiv preprint*, cs.CL/2303.08774v3.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. arXiv cs.CL/2203.02155.

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

Charlotte Pouw, Nora Hollenstein, and Lisa Beinborn. 2023. Cross-lingual transfer of cognitive processing complexity. In *Findings of EACL 2023*, pages 655–669.

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, 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.

Alfréd Rényi. 1961. 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*, volume 4.1, pages 547–562. University of California Press.

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

Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. 2022. Multitask prompted training enables zero-shot task generalization. In *ICLR 2022-Tenth International Conference on Learning Representations*.

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

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

C E Shannon. 1948. A Mathematical Theory of Communication. *Bell System Technical Journal*, 27(3):379–423.

Stuart C Shapiro. 2003. Artificial intelligence (AI). In *Encyclopedia of Computer Science*, pages 89–93. John Wiley and Sons Ltd., GBR.

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. *Journal of Cognition*, 128(3):302–319.

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

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

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford Alpaca: An instruction-following LLaMA model. `https://github.com/tatsu-lab/stanford_alpaca`.

William Timkey and Tal Linzen. 2023. A language model with limited memory capacity captures interference in human sentence processing. In *Findings of EMNLP 2023*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. LLaMA 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Marten van Schijndel and Tal Linzen. 2019. Can entropy explain successor surprisal effects in reading? In *Proceedings of SCiL 2019*, pages 1–7.

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

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

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

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

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of EMNLP 2020 : System Demonstrations*, pages 38–45.

Canwen Xu, Daya Guo, Nan Duan, and Julian McAuley. 2023. Baize: An open-source chat model with parameter-efficient tuning on self-chat data. In *Proceedings of EMNLP 2023*, pages 6268–6278.

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

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, and Guoyin Wang. 2023. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*.

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, Yuchen Eleanor Jiang, Ethan Wilcox, Ryan Cotterell, and Mrinmaya Sachan. 2023. Controlled text generation with natural language instructions. In *Proceedings of ICML 2023*, volume 202, pages 42602–42613.

A Models

We used the LLM implementations available via the huggingface transformer library (Wolf et al., 2020). The exact model URLs are listed in Table 7. Some large models are loaded using quantization. We used them for text decoding or probability computation; such usage of models follows their license and intended use. A single NVIDIA A100 GPU (40GB) was used for the experiments.

As of the date we experimented (2023/10/20), we can not get the probability information from recent OpenAI models such as GPT-4 (OpenAI, 2023). Thus, we only used legacy models that can yield probability scores. We also excluded the `gpt-3.5-turbo-instruct` model since the use of logprobs and echo options is restricted for this model to compute probabilities.

B PPL–PPP relationship with the explicit instruction to complete the sentence

Figure 4 shows the same experiments as §3.2 except for using the prompt: *Please complete the following sentence:*. The results were consistent with §3.2; the IT-LLMs exhibited competitive or worse PPP against the PPP–PPL relationship estimated by base LMs.

C Prompting

C.1 Prompt lists

Table 8 shows the prompts used in our experiments. We used the format-1 prompts for Falcon and GPT-3.5 models and the format-2 for LLaMA-2 models. That is, the format-2 is used in §4.1. Table 4 in §3 shows the averaged results using the respective formats. Figure 3 shows all the results, including LLaMA-2 with format-1 prompts and Falcon and GPT-3.5 with format-2 prompts.

Why? We observed that LLaMA-2 models tend to generate additional user inquiries when using the format-1 prompts; thus, by format-2 prompts, we made it clear for LLaMA-2 models to play the system roles and special instruction marking symbols [INST].[12] For example, LLaMA-2 completed the format-1 prompt (red words are generated) as follows: *Please complete the following sentence: \n They were playing on the _________ when they found the treasure. \n A. beach \n B. mountain \n C. park \n D. forest \n Answer: A. beach.* Conversely, with the format-1 prompt, Falcon and GPT-3.5 models are shown to complete the sentence intendedly and exhibit better PPLs than format-2 results; thus, we used the format-1 prompts for these LLMs.

C.2 Details on prompt biases

Setting details: We appended the first five words in the second sentence in DC's each document to a particular prompt for sampling sentences. When a model generated multiple sentences, the first one was used for the analysis in §4.1. We set top_p threshold to 0.95. In the analysis in §4.1, Zipf frequency was computed with the `word_freq` package (Speer, 2022); here, stopwords were excluded using nltk. Sentence and dependency lengths were computed using the nltk sentence/word tokenizers and spaCy dependency parser (`en_core_web_sm` version). Notably, some of the statistics (dependency length distribution) form non-normal distribution; reporting the average score (Table 3) would not be suitable. Nevertheless, other statistics, such as skewness, yielded similar results. Thus, we tentatively adopted the report of average values for simplicity. Tables 9 and 10 show the example of completed sentences conditioned by different prompts.

C.3 Detailed results

Tables 11, 12, 13, and 14 present the detailed results shown in §4. The advantage of prompts to use simple vocabulary/grammar demonstrated in §4 is generally reproduced across different models and corpora, but GPT 3.5 models sometimes prefer the instruction to "carefully focus on grammar" rather than use simple grammar.

D Metalinguistic prompting

Tables 15 and 16 show the exact prompts for directly asking LMs about the word's cognitive load or probability. Note that a token ID is attached with respect to the token position in a sentence to distinguish the multiple tokens sharing the same word in a sentence.

[12]Based on the LLaMA-2 format information: https://github.com/samrawal/llama2_chat_templater

Model	Instruction-tuning	Link	Quant.
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	

Table 7: Information about the LLMs.

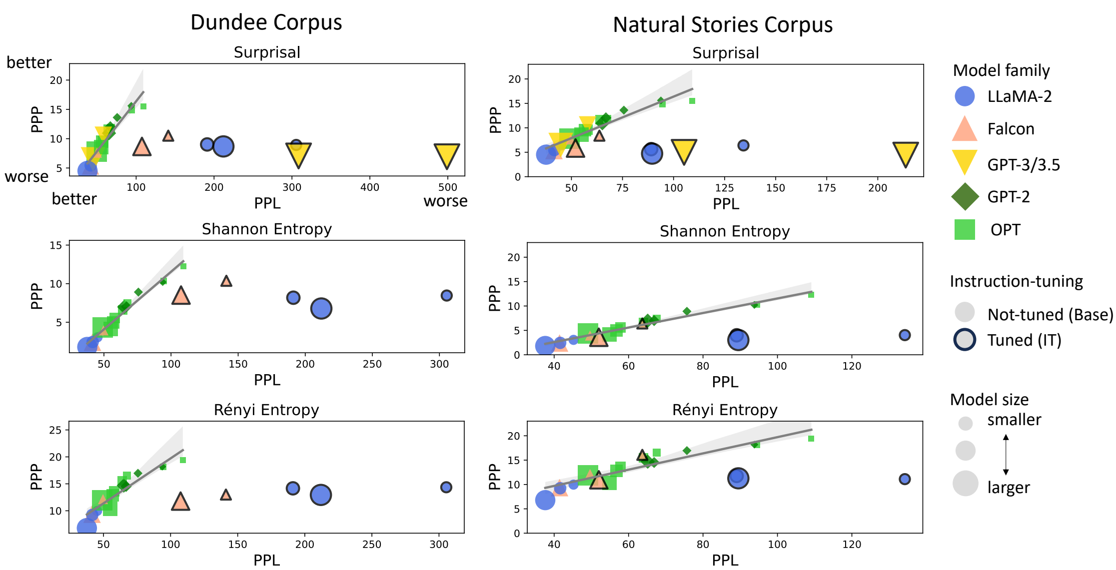

Figure 4: The relationship between PPL and PPP when using the prompt: *Please complete the following sentence:*. Each point corresponds to each LLM, and those with black-edged outlines are IT-LLMs. The PPP–PPL regression line is estimated by base LLMs, and the colored area presents a 95% confidence interval. IT-LLMs were relatively poor (below the line) at balancing PPL and PPP.

Format 1 (GPT3.5, Falcon)	Format 2 (LLaMA-2)
Please complete the following sentence to make it as grammatically simple as possible: \n $w_0, \cdots, w_{t-1}$	[INST] Please generate a grammatically simple sentence as much as possible. [/INST] Answer: $w_0, \cdots, w_{t-1}$
Please complete the following sentence with a careful focus on grammar \n $w_0, \cdots, w_{t-1}$	[INST] Please generate a sentence with a careful focus on grammar. [/INST] Answer: $w_0, \cdots, w_{t-1}$
Please complete the following sentence to make it as grammatically complex as possible: \n $w_0, \cdots, w_{t-1}$	[INST] Please generate a grammatically complex sentence as much as possible. [/INST] Answer: $w_0, \cdots, w_{t-1}$
Please complete the following sentence using the simplest vocabulary possible: \n $w_0, \cdots, w_{t-1}$	[INST] Please generate a sentence using the simplest vocabulary possible. [/INST] Answer: \n $w_0, \cdots, w_{t-1}$
Please complete the following sentence with a careful focus on word choice \n $w_0, \cdots, w_{t-1}$	[INST] Please generate a sentence with a careful focus on word choice. [/INST] Answer: \n $w_0, \cdots, w_{t-1}$
Please complete the following sentence using the most difficult vocabulary possible: \n $w_0, \cdots, w_{t-1}$	[INST] Please generate a sentence using the most difficult vocabulary possible. [/INST] Answer: \n $w_0, \cdots, w_{t-1}$
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 $w_0, \cdots, w_{t-1}$	[INST] 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. [/INST] Answer: \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 $w_0, \cdots, w_{t-1}$	[INST] 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. [/INST] Answer: \n $w_0, \cdots, w_{t-1}$
[INST] Please complete the following sentence: \n $w_0, \cdots, w_{t-1}$	Please generate a sentence. [/INST] Answer: \n $w_0, \cdots, w_{t-1}$

Table 8: All the prompts we used in our experiments.

Please generate a grammatically simple sentence as much as possible. Answer: At that time he said, "I am going to the store."
Please generate a sentence with a careful focus on grammar. Answer: At that time he said he was going to the store, but he went to the park instead.
Please generate a grammatically complex sentence as much as possible. Answer: At that time he said that he had seen her not more than once or twice since they had last met, when he had been so struck by her beauty that he had almost lost his breath, and he had not been able to forget her ever since, despite his many attempts
Please generate a sentence using the simplest vocabulary possible. Answer: At that time he said, "I can't go."
Please generate a sentence with a careful focus on word choice. Answer: At that time he said to himself, "I'm going to go out and do something that will make me feel really good."
Please generate a sentence using the most difficult vocabulary possible. Answer: At that time he said to himself, "I must needs make an inscrutable and enigmatic allusion to the pervasive and pernicious effects of the miasmic and noxious atmosphere which permeates and pervades this putative
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. Answer: At that time he said he would be able to predict the next words but he was wrong.
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. Answer: At that time he said he was going to the store, but he actually went to the library.
Please generate a sentence. Answer: At that time he said to me, "My dear friend, I must go to the store."

Table 9: Examples of the completions by LLaMA-2 70B with different prompts.

Please generate a grammatically simple sentence as much as possible. Answer: The evidence of the annual rings is the most important proof that the tree grows.
Please generate a sentence with a careful focus on grammar. Answer: The evidence of the annual festival was apparent in the colorful decorations adorning the streets.
Please generate a grammatically complex sentence as much as possible. Answer: The evidence of the annual festival's increasing popularity and cultural significance, as demonstrated by the growing number of attendees and the expanding variety of musical performances, food vendors, and artisanal goods on display, has made it imperative for the organiz
Please generate a sentence using the simplest vocabulary possible. Answer: The evidence of the annual flooding was obvious.
Please generate a sentence with a careful focus on word choice. Answer: The evidence of the annual festival was scattered across the grounds, with discarded flyers and abandoned souvenirs littering the once-pristine lawns.
Please generate a sentence using the most difficult vocabulary possible. Answer: The evidence of the annual financial audit conducted by the external accounting firm indicated that the company's fiscal year-end consolidated balance sheet contained several glaring discrepancies, which necessitated a comprehensive reconciliation process to rectify the
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. Answer: The evidence of the annual performance review was scattered all over the table, making it difficult for the manager to piece together a coherent picture of the employee's strengths and weaknesses.
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. Answer: The evidence of the annual performance review was overwhelming.
Please generate a sentence. Answer: The evidence of the annual rings in the cross section of a tree trunk provides valuable information to dendrochronologists.

Table 10: Examples of the completions by LLaMA-2 70B with different prompts.

Prompt	Model	h	H	$H_{0.5}$	PPL
Please complete the following sentence to make it as grammatically simple as possible	Llama-2 7B	9.10	9.15	15.58	341.90
	Llama-2 13B	9.21	8.37	15.18	184.71
	Llama-2 70B	9.03	6.63	12.82	139.97
	Falcon 7B	10.54	10.60	13.20	153.40
	Falcon 40B	8.91	8.93	12.18	106.49
	GPT-3.5 D2	6.20	-	-	524.09
	GPT-3.5 D3	6.20	-	-	429.86
Please complete the following sentence with a careful focus on grammar	Llama-2 7B	9.29	9.07	15.15	250.90
	Llama-2 13B	9.36	7.85	14.51	169.02
	Llama-2 70B	9.25	6.62	12.71	152.12
	Falcon 7B	10.02	10.48	13.10	168.28
	Falcon 40B	8.90	8.55	11.70	103.18
	GPT-3.5 D2	6.51	-	-	620.67
	GPT-3.5 D3	6.61	-	-	331.92
Please complete the following sentence to make it as grammatically complex as possible	Llama-2 7B	9.12	8.28	14.30	315.23
	Llama-2 13B	8.20	7.39	13.87	241.85
	Llama-2 70B	8.35	5.73	12.02	191.69
	Falcon 7B	10.25	10.39	12.98	143.58
	Falcon 40B	8.67	8.80	12.15	105.21
	GPT-3.5 D2	5.81	-	-	1.014.37
	GPT-3.5 D3	6.00	-	-	458.79
Please complete the following sentence using the simplest vocabulary possible	Llama-2 7B	9.01	9.27	15.45	362.27
	Llama-2 13B	8.68	8.63	15.29	300.47
	Llama-2 70B	8.60	6.52	13.07	214.54
	Falcon 7B	10.41	10.77	13.09	167.87
	Falcon 40B	8.57	8.72	12.16	115.59
	GPT-3.5 D2	5.89	-	-	638.71
	GPT-3.5 D3	5.96	-	-	630.27
Please complete the following sentence with a careful focus on word choice	Llama-2 7B	9.10	8.58	14.72	269.79
	Llama-2 13B	8.98	7.55	14.05	175.18
	Llama-2 70B	8.80	6.42	12.51	168.91
	Falcon 7B	9.64	10.06	12.90	166.03
	Falcon 40B	8.67	8.30	11.65	105.26
	GPT-3.5 D2	6.73	-	-	453.51
	GPT-3.5 D3	6.53	-	-	329.08
Please complete the following sentence using the most difficult vocabulary possible	Llama-2 7B	7.68	7.77	13.28	515.02
	Llama-2 13B	7.23	7.42	13.34	376.41
	Llama-2 70B	7.82	5.82	12.08	275.86
	Falcon 7B	9.63	10.14	12.76	164.03
	Falcon 40B	7.99	8.44	11.91	112.07
	GPT-3.5 D2	5.56	-	-	536.87
	GPT-3.5 D3	5.98	-	-	348.27
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	Llama-2 7B	9.88	8.96	14.79	223.11
	Llama-2 13B	9.85	8.40	14.70	170.97
	Llama-2 70B	9.69	6.86	13.17	175.96
	Falcon 7B	9.20	9.20	12.73	230.78
	Falcon 40B	8.45	8.52	11.96	133.90
	GPT-3.5 D2	5.74	-	-	1.155.36
	GPT-3.5 D3	5.34	-	-	612.23
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	Llama-2 7B	9.88	8.96	14.61	212.73
	Llama-2 13B	9.38	8.43	14.44	192.83
	Llama-2 70B	9.70	6.84	12.95	166.50
	Falcon 7B	9.57	9.99	13.04	219.45
	Falcon 40B	8.76	8.27	11.79	121.70
	GPT-3.5 D2	6.26	-	-	1.085.46
	GPT-3.5 D3	5.65	-	-	613.11
Please complete the following sentence	Llama-2 7B	8.71	8.63	14.86	307.29
	Llama-2 13B	8.74	8.16	14.52	191.69
	Llama-2 70B	8.49	6.35	12.62	193.78
	Falcon 7B	10.52	10.37	13.01	141.23
	Falcon 40B	8.68	8.52	11.80	107.46
	GPT-3.5 D2	6.91	-	-	498.89
	GPT-3.5 D3	7.05	-	-	308.50

Table 11: Full results of the experiments in §4 on the DC with the format-1 prompts.

Prompt	Model	h	H	$\mathrm{H}_{0.5}$	PPL
[INST] Please generate a grammatically simple sentence as much as possible. [/INST] Answer:	LLaMA-2 7B	8.88	7.07	13.72	325.44
	LLaMA-2 13B	8.35	6.08	12.58	373.04
	LLaMA-2 70B	7.38	4.67	10.59	466.80
[INST] Please generate a sentence with a careful focus on grammar. [/INST] Answer:	LLaMA-2 7B	9.25	6.79	13.48	232.50
	LLaMA-2 13B	8.59	5.75	12.21	307.64
	LLaMA-2 70B	8.05	4.15	10.13	345.80
[INST] Please generate a grammatically complex sentence as much as possible. [/INST] Answer:	LLaMA-2 7B	8.59	6.32	12.41	312.61
	LLaMA-2 13B	8.35	5.68	11.94	329.54
	LLaMA-2 70B	7.58	4.92	10.63	396.32
[INST] Please generate a sentence using the simplest vocabulary possible. [/INST] Answer:	LLaMA-2 7B	9.02	7.80	13.84	352.90
	LLaMA-2 13B	7.55	6.34	12.75	564.19
	LLaMA-2 70B	7.21	4.84	10.57	557.85
[INST] Please generate a sentence with a careful focus on word choice. [/INST] Answer:	LLaMA-2 7B	8.56	5.84	12.30	267.50
	LLaMA-2 13B	8.00	5.63	11.85	327.15
	LLaMA-2 70B	7.75	4.07	9.86	337.14
[INST] Please generate a sentence using the most difficult vocabulary possible. [/INST] Answer:	LLaMA-2 7B	8.22	5.93	12.05	351.65
	LLaMA-2 13B	7.85	5.67	11.99	345.42
	LLaMA-2 70B	7.28	4.43	10.29	597.32
[INST] 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. [/INST] Answer:	LLaMA-2 7B	8.97	7.19	13.46	317.99
	LLaMA-2 13B	8.86	6.23	12.65	288.86
	LLaMA-2 70B	8.65	4.56	11.05	364.23
[INST] 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. [/INST] Answer:	LLaMA-2 7B	9.15	7.95	14.33	292.45
	LLaMA-2 13B	9.26	6.77	13.47	247.37
	LLaMA-2 70B	8.30	4.86	11.52	318.45
[INST] Please generate a sentence. [/INST] Answer:	LLaMA-2 7B	9.08	6.61	13.34	223.67
	LLaMA-2 13B	8.33	5.66	12.28	292.13
	LLaMA-2 70B	7.85	4.68	10.84	298.61

Table 12: Full results of the experiments in §4 on the DC with the format-2 prompts.

Prompt	Model	h	H	$H_{0.5}$	PPL
Please complete the following sentence to make it as grammatically simple as possible	Llama-2 7B	7.60	3.88	11.22	149.88
	Llama-2 13B	6.42	3.04	10.68	84.31
	Llama-2 70B	5.63	1.19	7.61	64.20
	Falcon 7B	8.74	6.20	16.48	69.38
	Falcon 40B	6.03	2.99	10.11	50.95
	GPT-3.5 D2	4.97	-	-	201.88
	GPT-3.5 D3	6.69	-	-	155.32
Please complete the following sentence with a careful focus on grammar	Llama-2 7B	6.48	3.61	10.64	117.67
	Llama-2 13B	5.32	3.20	10.39	84.06
	Llama-2 70B	4.88	2.29	9.65	74.86
	Falcon 7B	9.05	6.32	16.20	73.42
	Falcon 40B	5.87	2.93	9.82	50.45
	GPT-3.5 D2	4.57	-	-	238.15
	GPT-3.5 D3	5.88	-	-	106.68
Please complete the following sentence to make it as grammatically complex as possible	Llama-2 7B	6.69	4.14	10.32	118.23
	Llama-2 13B	5.72	3.46	10.54	89.93
	Llama-2 70B	4.79	1.54	7.73	75.42
	Falcon 7B	8.51	5.36	15.12	62.39
	Falcon 40B	5.69	2.82	9.58	46.98
	GPT-3.5 D2	3.64	-	-	353.98
	GPT-3.5 D3	4.49	-	-	149.71
Please complete the following sentence using the simplest vocabulary possible	Llama-2 7B	6.74	4.18	11.88	158.02
	Llama-2 13B	5.48	5.46	13.99	133.04
	Llama-2 70B	4.83	2.93	10.67	94.71
	Falcon 7B	7.92	6.51	16.49	69.94
	Falcon 40B	5.62	2.97	10.03	53.52
	GPT-3.5 D2	4.66	-	-	245.26
	GPT-3.5 D3	4.92	-	-	214.47
Please complete the following sentence with a careful focus on word choice	Llama-2 7B	5.98	3.50	1-	127.13
	Llama-2 13B	4.79	3.32	10.52	89.39
	Llama-2 70B	4.74	2.71	10.29	82.48
	Falcon 7B	8.78	5.53	15.58	71.02
	Falcon 40B	5.82	2.89	9.93	50.66
	GPT-3.5 D2	4.92	-	-	174.62
	GPT-3.5 D3	5.67	-	-	107.04
Please complete the following sentence using the most difficult vocabulary possible	Llama-2 7B	4.61	3.48	9.45	193.86
	Llama-2 13B	4.44	3.26	9.83	148.45
	Llama-2 70B	3.62	2.90	9.70	124.40
	Falcon 7B	7.33	5.08	14.56	67.20
	Falcon 40B	4.68	2.21	8.50	53.00
	GPT-3.5 D2	4.03	-	-	172.35
	GPT-3.5 D3	3.74	-	-	110.72
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	Llama-2 7B	5.92	3.53	10.15	102.16
	Llama-2 13B	5.16	3.64	10.76	80.88
	Llama-2 70B	4.44	3.01	11.42	88.02
	Falcon 7B	6.26	5.27	15.28	90.87
	Falcon 40B	4.66	4.19	11.56	60.71
	GPT-3.5 D2	3.21	-	-	311.83
	GPT-3.5 D3	2.68	-	-	168.23
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	Llama-2 7B	6.10	3.60	10.11	95.46
	Llama-2 13B	4.95	3.83	10.83	81.96
	Llama-2 70B	4.42	3.19	10.90	82.45
	Falcon 7B	6.98	6.40	15.73	83.36
	Falcon 40B	5.01	3.87	11.35	53.71
	GPT-3.5 D2	3.49	-	-	304.80
	GPT-3.5 D3	2.97	-	-	173.41
Please complete the following sentence	Llama-2 7B	6.48	3.84	10.90	146.86
	Llama-2 13B	5.32	3.60	11.34	95.65
	Llama-2 70B	4.33	3.34	11.09	94.19
	Falcon 7B	8.39	6.42	16.09	63.69
	Falcon 40B	5.86	3.60	10.96	52.01
	GPT-3.5 D2	4.40	-	-	213.67
	GPT-3.5 D3	4.97	-	-	105.16

Table 13: Full results of the experiments in §4 on the NS with the format-1 prompts.

Prompt	Model	h	H	$H_{0.5}$	PPL
[INST] Please generate a grammatically simple sentence as much as possible. [/INST] Answer:	LLaMA-2 7B	4.60	3.04	9.35	146.17
	LLaMA-2 13B	3.70	3.10	10.44	186.45
	LLaMA-2 70B	3.22	1.57	6.55	193.25
[INST] Please generate a sentence with a careful focus on grammar. [/INST] Answer:	LLaMA-2 7B	5.04	2.59	8.44	100.95
	LLaMA-2 13B	3.88	2.19	7.92	128.30
	LLaMA-2 70B	4.09	0.87	5.06	140.79
[INST] Please generate a grammatically complex sentence as much as possible. [/INST] Answer:	LLaMA-2 7B	4.95	1.90	6.97	113.33
	LLaMA-2 13B	4.72	1.28	7.07	115.80
	LLaMA-2 70B	4.23	0.75	5.10	130.95
[INST] Please generate a sentence using the simplest vocabulary possible. [/INST] Answer:	LLaMA-2 7B	4.50	2.06	9.19	166.33
	LLaMA-2 13B	3.39	2.32	9.64	270.99
	LLaMA-2 70B	2.95	1.63	6.19	233.22
[INST] Please generate a sentence with a careful focus on word choice. [/INST] Answer:	LLaMA-2 7B	4.04	1.91	7.29	115.81
	LLaMA-2 13B	3.75	1.52	6.74	146.38
	LLaMA-2 70B	3.59	0.68	5.04	138.90
[INST] Please generate a sentence using the most difficult vocabulary possible. [/INST] Answer:	LLaMA-2 7B	4.32	1.21	6.25	151.10
	LLaMA-2 13B	4.04	1.00	6.02	143.37
	LLaMA-2 70B	2.62	1.14	5.74	221.19
[INST] 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. [/INST] Answer:	LLaMA-2 7B	4.24	3.54	10.99	125.88
	LLaMA-2 13B	4.57	2.90	10.93	117.23
	LLaMA-2 70B	4.08	1.38	7.00	143.37
[INST] 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. [/INST] Answer:	LLaMA-2 7B	5.22	3.34	10.49	118.79
	LLaMA-2 13B	4.28	3.11	10.63	106.47
	LLaMA-2 70B	4.79	1.46	7.17	127.61
[INST] Please generate a sentence. [/INST] Answer:	LLaMA-2 7B	5.38	2.98	8.91	91.26
	LLaMA-2 13B	5.10	1.73	8.74	115.99
	LLaMA-2 70B	4.33	1.12	5.82	121.74

Table 14: Full results of the experiments in §4 on the NS with the format-2 prompts.

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,

Suppose humans read the following sentence: "A clear and joyous day it was and out on the wide open sea, thousands upon thousands of sparkling water drops, excited by getting to play in the ocean, danced all around."
List the tokens and their IDs in order of their reading cost (high to low) during sentence processing.
Token ID:
0: A, 1: clear, 2: and, 3: joyous, 4: day, 5: it, 6: was, 7: and, 8: out, 9: on, 10: the, 11: wide, 12: open, 13: sea,, 14: thousands, 15: upon, 16: thousands, 17: of, 18: sparkling, 19: water, 20: drops,, 21: excited, 22: by, 23: getting, 24: to, 25: play, 26: in, 27: the, 28: ocean,, 29: danced, 30: all, 31: around.,
Answer:
13: sea,, 20: drops,, 28: ocean,, 21: excited, 0: A, 2: and, 22: by, 12: open, 7: and, 31: around., 19: water, 27: the, 3: joyous, 29: danced, 6: was, 23: getting, 11: wide, 18: sparkling, 30: all, 17: of, 14: thousands, 24: to, 15: upon, 4: day, 25: play, 1: clear, 8: out, 16: thousands, 5: it, 26: in, 9: on, 10: the,

Suppose humans read the following sentence: "By the handsome reward many felt tempted, but the thought of the boar with its deadly tusks and face like thunder soon put an end to their ambitions."
List the tokens and their IDs in order of their reading cost (high to low) during sentence processing.
Token ID:
0: By, 1: the, 2: handsome, 3: reward, 4: many, 5: felt, 6: tempted,, 7: but, 8: the, 9: thought, 10: of, 11: the, 12: boar, 13: with, 14: its, 15: deadly, 16: tusks, 17: and, 18: face, 19: like, 20: thunder, 21: soon, 22: put, 23: an, 24: end, 25: to, 26: their, 27: ambitions.,
Answer:
4: many, 27: ambitions., 3: reward, 5: felt, 2: handsome, 8: the, 6: tempted,, 26: their, 7: but, 21: soon, 1: the, 16: tusks, 12: boar, 19: like, 20: thunder, 13: with, 17: and, 14: its, 15: deadly, 22: put, 0: By, 10: of, 11: the, 18: face, 9: thought, 23: an, 24: end, 25: to,

Suppose humans read the following sentence: <TARGET SENT>
List the tokens and their IDs in order of their reading cost (high to low) during sentence processing.
Token ID:
<TOKENS FROM TARGET SENT>
Answer:

Table 15: An example of a prompt for asking about the processing cost of words.

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,

Suppose you read the following sentence: "A clear and joyous day it was and out on the wide open sea, thousands upon thousands of sparkling water drops, excited by getting to play in the ocean, danced all around."
List the tokens and their IDs in order of their probability in context (low to high).
Token ID:
0: A, 1: clear, 2: and, 3: joyous, 4: day, 5: it, 6: was, 7: and, 8: out, 9: on, 10: the, 11: wide, 12: open, 13: sea,, 14: thousands, 15: upon, 16: thousands, 17: of, 18: sparkling, 19: water, 20: drops,, 21: excited, 22: by, 23: getting, 24: to, 25: play, 26: in, 27: the, 28: ocean,, 29: danced, 30: all, 31: around.,
Answer:
3: joyous, 21: excited, 14: thousands, 23: getting, 0: A, 8: out, 18: sparkling, 20: drops,, 1: clear, 5: it, 11: wide, 19: water, 30: all, 7: and, 15: upon, 28: ocean,, 29: danced, 13: sea,, 4: day, 9: on, 25: play, 31: around., 22: by, 24: to, 12: open, 2: and, 26: in, 6: was, 27: the, 10: the, 17: of, 16: thousands,

Suppose you read the following sentence: "By the handsome reward many felt tempted, but the thought of the boar with its deadly tusks and face like thunder soon put an end to their ambitions."
List the tokens and their IDs in order of their probability in context (low to high).
Token ID:
0: By, 1: the, 2: handsome, 3: reward, 4: many, 5: felt, 6: tempted,, 7: but, 8: the, 9: thought, 10: of, 11: the, 12: boar, 13: with, 14: its, 15: deadly, 16: tusks, 17: and, 18: face, 19: like, 20: thunder, 21: soon, 22: put, 23: an, 24: end, 25: to, 26: their, 27: ambitions.,
Answer:
2: handsome, 3: reward, 12: boar, 4: many, 18: face, 5: felt, 0: By, 6: tempted,, 21: soon, 9: thought, 20: thunder, 13: with, 15: deadly, 27: ambitions., 23: an, 7: but, 19: like, 1: the, 8: the, 17: and, 26: their, 11: the, 14: its, 22: put, 16: tusks, 10: of, 24: end, 25: to,

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:

Table 16: An example of a prompt for asking about the word probability.