What Makes Language Models Good-enough?

Daiki Asami
University of Delaware
daiasami@udel.edu

Saku Sugawara
National Institute of Informatics
saku@nii.ac.jp

Abstract

Psycholinguistic research suggests that humans may build a representation of linguistic input that is 'good-enough' for the task at hand. This study examines what architectural features make language models learn human-like good-enough language processing. We focus on the number of layers and self-attention heads in Transformers. We create a good-enough language processing (GELP) evaluation dataset (7,680 examples), which is designed to test the effects of two plausibility types, eight construction types, and three degrees of memory cost on language processing. To annotate GELP, we first conduct a crowdsourcing experiment whose design follows prior psycholinguistic studies. Our model evaluation against the annotated GELP then reveals that the full model as well as models with fewer layers and/or self-attention heads exhibit a good-enough performance. This result suggests that models with shallower depth and fewer heads can learn good-enough language processing.[1]

1 Introduction

Language models exhibit impressive performance in various natural language understanding tasks (Devlin et al., 2019; Brown et al., 2020; Mahowald et al., 2023), but one common concern is that they often rely on heuristics (Geirhos et al., 2020; Du et al., 2023). For instance, BERT (Devlin et al., 2019) makes predictions based on surface features, which leads to poor results in adversarial examples (McCoy et al., 2019). Large language models such as GPT-2 (Radford et al., 2019) also adopt fallible heuristics in in-context learning (Tang et al., 2023).

However, it is too hasty to view models' reliance on heuristics as a flaw. According to a 'good-enough' theory of human sentence processing (Ferreira, 2003; Ferreira and Patson, 2007; Christianson, 2016), humans also adopt some types of heuristics.[2] For instance, they may build an incomplete representation of linguistic input; as a result, they occasionally misinterpret an impossible description (e.g., *The ball kicked the boy*; Ferreira, 2003). The good-enough theory posits that such fallible language processing is good-enough for everyday communication. In addition, it allows humans to efficiently process linguistic input by saving cognitive resources as suggested by the findings that they tend to rely on it when they face cognitive demands (e.g., (1) vs. (2) in Figure 1; Christianson et al., 2001, 2006, 2010; Patson et al., 2006).

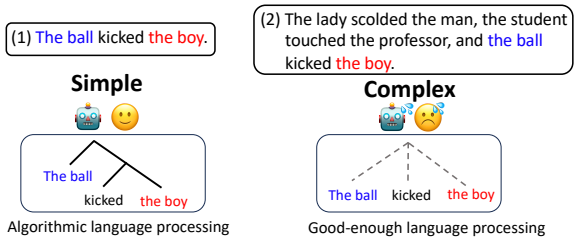


Figure 1: While algorithmic language processing involves a detailed syntactic analysis, good-enough language processing involves an incomplete one. We hypothesize that humans are more likely to adopt a good-enough strategy when processing a complex sentence than a simple one. A human-like good-enough model should exhibit a similar tendency.

In light of this cognitive background, we view language models' adaptation of heuristics as their potential for human-like linguistic performance (Linzen, 2020; Hagendorff and Fabi, 2023). An open question here is what architectural features make them learn human-like good-enough language processing. To study this question, we explore how the numbers of layers and self-attention heads affect models' performance. Prior studies suggest that the model depth and attention heads

[1] Our dataset and codebase for dataset creation are available at https://github.com/nii-cl/gelp.

[2] Besides 'good-enough', the psycholinguistic literature cited in the main text uses terms such as 'heuristic', 'shallow', and 'underspecified'. We use 'good-enough' unless the difference is crucial.

Findings of the Association for Computational Linguistics ACL 2024, pages 15453–15467

Construction	Implausible premise	Hypothesis (correct label)
(a) Transitive	The ball kicked the boy.	The boy kicked the ball. (N) The ball kicked the boy. (E)
(b) Passive	The boy was kicked by the ball.	The ball was kicked by the boy. (N) The boy was kicked by the ball. (E)
(c) DOC	The boy gave the apple the girl.	The boy gave the girl the apple. (N) The boy gave the apple the girl. (E)
(d) Dative	The boy gave the girl to the apple.	The boy gave the apple to the girl. (N) The boy gave the girl to the apple. (E)
(e) Ben. DOC	The cook made the bread the man.	The cook made the man the bread. (N) The cook made the bread the man. (E)
(f) Ben. *for*	The cook made the man for the bread.	The cook made the bread for the man. (N) The cook made the man for the bread? (E)
(g) Exp. Subj.	The book liked the girl.	The book liked the girl. (N) The girl liked the book. (E)
(h) Exp. Obj.	The girl pleased the book.	The girl pleased the book. (N) The book pleased the girl. (E)

Table 1: Eight constructions in GELP with example implausible premises and corresponding hypotheses. Abbreviations: Ben. = Benefactive; DOC = double object construction; Exp. = Experiencer; Obj. = Object; Subj. = Subject.

Memory load	Examples
Low (one proposition)	The ball kicked the boy.
Medium (two propositions)	The girl bought the cup and the ball kicked the boy.
High (three propositions)	The girl bought the cup, the singer broke the window, and the ball kicked the boy.

Table 2: Three degrees of memory load. The low, medium, and high memory load conditions include one, two, and three propositions, respectively.

contribute to syntactic generalizations (Mueller and Linzen, 2023) and parallel a human working memory system (Ryu and Lewis, 2021; Timkey and Linzen, 2023), respectively. Considering these findings, we hypothesize that (i) our aimed good-enough model requires a shallow depth because it does not have to engage in a detailed syntactic analysis and (ii) needs a small number of heads because a strong memory system is not necessary.

To test these hypotheses, we evaluate BERT's language processing capabilities through the lens of misinterpretation of sentences. We create a good-enough language processing (GELP) dataset with 7,680 items. GELP includes not only plausible but also implausible items to investigate humans' as well as models' misinterpretation. Additionally, considering the prior psycholinguistic finding that some constructions are more likely to cause misinterpretation than others (Gibson et al., 2013), it targets eight types of constructions (Table 1). Finally, to examine the effect of the memory demand, it operationalizes memory cost by including items with one, two, or three propositions (Table 2).

We first annotate the GELP dataset and probe humans' language processing by conducting a crowdsourcing experiment whose design follows previous psycholinguistic studies (Christianson et al., 2001, 2006). Against the human data, we test which model with different architectures behaves in a human-like fashion. We find that among 24 variants of BERT, those with fewer layers and heads perform in a good-enough way, similar to the full BERT-base model. This result suggests that a deep architecture is not necessary for a model to learn good-enough language processing, which is consistent with our hypothesis (i). In contrast, a closer look indicates that the role of attention heads in good-enough language processing does not confirm our hypothesis (ii). This study modestly informs psycholinguistics by supporting the claim that humans' fallible language processing stems from a shallow syntactic analysis and suggesting that it has to do with the encoding phase of the working memory.

2 Background and Motivation

2.1 Good-enough Theory in Psycholinguistics

A good-enough theory of language processing posits that humans may build representations that are good enough to achieve their communicative goal (Ferreira, 2003; Ferreira and Patson, 2007; Christianson, 2016). Such language processing is fallible in language use. For instance, Ferreira

(2003) find that humans misinterpret an implausible sentence (e.g., *The professor bit the dog*) as its plausible version (e.g., *The dog bit the professor*). They reason that such misinterpretation results from canonical word order and plausibility heuristics. Specifically, the plausible sentence is consistent with canonical patterns of English (i.e., a noun–verb–noun order generally represents an agent–action–patient relation) and world knowledge (i.e., dogs usually bite people, but not vice versa). Another interpretation of humans' misinterpretation is that they may build a shallow representation of linguistic input (Sanford and Sturt, 2002; Sanford and Graesser, 2006). Gibson et al. (2013) add to Ferreira (2003)'s finding by showing that the misinterpretation of implausible sentences depends on types of constructions.[3]

Although good-enough language processing can cause errors, it enables humans to process language efficiently. Previous studies suggest that such language processing is likely to take place under cognitively demanding tasks such as a priming paradigm involving both language comprehension and production (Christianson et al., 2010) and a reading experiment with structurally complex stimuli (Christianson et al., 2001; Ferreira, 2003). Additionally, other studies indicate that a limited working memory capacity can motivate humans to rely on a good-enough strategy (Christianson et al., 2006; Patson et al., 2006). All of these findings are consistent with the view that humans adopt fallible but efficient good-enough language processing to reduce cognitive costs.

Crucially, humans can build a detailed representation if necessary. For instance, misinterpretation does not easily occur in a situation that requires deep processing such as proofreading. Ferreira (2003) emphasizes that humans' language processing has a good balance between robust, algorithmic language processing and non-robust, heuristic language processing.

2.2 Language Models' Reliance on Heuristics

Heuristics also receive much attention in research on natural language understanding by language models (Geirhos et al., 2020; Du et al., 2023). Language models are known to learn various types of heuristics based on training data during fine-tuning in natural language inference (McCoy et al., 2019; Gururangan et al., 2018), question answering (Jia and Liang, 2017; Sugawara et al., 2018; Lai et al., 2021), and coreference inference (Zhao et al., 2018) tasks. Additionally, Tang et al. (2023) find that non-fine-tuned large language models such as GPT-2 (Radford et al., 2019) also adopt heuristics in in-context learning, suggesting that the recent models still suffer from the heuristics.

Despite this large body of research, it is an open question to what extent language models' reliance on heuristics resembles humans'. This question is crucial because implementing human-like heuristics into language models can lead to a more efficient system in terms of resource and computational requirements (Hagendorff and Fabi, 2023).

To tackle this question, we explore what architectural features contribute to the aimed model by focusing on the numbers of layers and self-attention heads in Transformers (Vaswani et al., 2017). We target these two architectural features because of their putative resemblance to human linguistic as well as non-linguistic cognitive systems. Mueller and Linzen (2023) show that model depth is important for models to learn syntactic generalizations. Assuming the psycholinguistic claim that good-enough language processing does not involve a detailed syntactic analysis (Sanford and Sturt, 2002; Sanford and Graesser, 2006; Ferreira, 2003; Ferreira and Patson, 2007; Christianson, 2016), our first hypothesis (i) is that our aimed model does not require a deep architecture. Regarding the self-attention heads, Ryu and Lewis (2021) and Timkey and Linzen (2023) suggest that a self-attention mechanism exhibits similarity to the retrieval phase of the working memory, but it remains open whether it also captures other phases (i.e., encoding and maintenance). Given that adaptation of heuristics may result from demand on the working memory system (Christianson et al., 2006; Patson et al., 2006), our second hypothesis (ii) is that models with fewer heads behave in a human-like fashion, capturing humans' reliance on a good-enough strategy as a function of the memory demand.

3 Dataset Creation

To probe models' language processing, we create a natural language inference (NLI) dataset called GELP, targeting two plausibility types, eight types of constructions, and three degrees of memory load.

[3] Gibson et al. (2013) do not adopt the good-enough theory but we introduce them because their findings are relevant to our study regardless of the theoretical framework that they adopt.

In an NLI task, models predict whether a sentence (*premise*) entails, contradicts, or is neutral to another (*hypothesis*; Condoravdi et al., 2003; Bowman et al., 2015). For the human–model comparison discussed in Section 5.2, our labels include *entailment* and *non-entailment*, the latter of which covers both *contradiction* and *neutral*.

3.1 Low Memory Load Condition

We first make items in the low memory load condition. Considering the effect of construction types on misinterpretation discussed in Section 2.1 (Gibson et al., 2013), GELP targets eight constructions (e.g., (a-h) in Table 1). We select 40 verbs from Levin (1993) for (1) transitive/passive, (2) double object/dative, (3) benefactive double object/*for*, (4) experiencer-subject, and (5) experiencer-object constructions each (a total of 200 verbs).[4] These verbs take both animate and inanimate arguments within a single sentence, which allows us to make implausible sentences by swapping them.

With these verbs, we create 80 plausible premises for each construction (a total of 640 contexts) by giving GPT-3.5-turbo[5] prompts.[6] We instruct it to make sentences with our selected verbs by using animate and inanimate nouns in positions of interest, which this paper indicates with red and blue for animate and inanimate nouns, respectively (e.g., *The boy kicked the ball*). We manually check all generated sentences and correct any noticeable errors by hand (e.g., if an animate noun appears in an object position of *kick*, we change it into an inanimate noun such as *ball*). Then, we swap the animate and inanimate nouns in each premise, creating 640 implausible premises (e.g., *The ball kicked the boy*). Finally, we make two hypotheses with *entailment* and *non-entailment* labels for each premise. As a result, the low memory condition has 2,560 pairs ({8 constructions}*{80 premises}*{2 plausibility types}*{2 hypothesis types}).

3.2 Medium and High Memory Load Conditions

Using the items in the low memory condition, we make pairs in the medium and high memory load conditions. In doing so, we use templates, which allow us to create a large number of items systematically. The medium memory load condition has templates for a premise and hypothesis such as (1).

(1) a. Target and the N1 V1 the N2.
b. Entailed Hypothesis

The premise consists of two propositions coordinated by one of the five connectives (*and*, *after*, *when*, *but*, and *because*). One proposition is a *target* sentence (i.e., Target) that corresponds to a premise in the low memory condition, and the other is a template-generated *filler* sentence (i.e., *the N1 V1 the N2*). For N1, V1, and V2 in the template, we use 201 transitive verbs and 515 animate nouns from Fedorenko et al. (2020). Every selected noun is a plausible subject or object of every selected verb. We ensure that these lexical items have no overlap with those used in target premises. The target sentence either proceeds or follows the filler (e.g., *Target and Filler* or *Filler and Target*) to prevent one from developing a strategy to identify which proposition they should focus on while ignoring the other one. A hypothesis can be either *Entailed Hypothesis* or *Non-entailed Hypothesis*, which correspond to hypotheses with *entailment* and *non-entailment* labels in the low memory condition, respectively. Combining the five connectives, two target-filler premise orders, and two hypothesis types results in 80 templates.

A premise in the high memory load condition consists of three propositions coordinated by two of the five connectives used in the medium memory load condition. We provide an example template for a premise in (2).

(2) Target and the N1 V1 the N2 but the N3 V2 the N4

For a hypothesis, we use the same template as in the medium memory condition. There are 20 permutation patterns of the two connectives out of the five connectives. A target sentence appears in one of the three positions within the premise. Combining the 20 connective patterns with the three proposition orders results in 60 templates.

Using the templates, we generate 2,560 pairs in the medium and high memory conditions each. Consequently, GELP has a total of 7,680 items (2,560 pairs in three memory load conditions each).

4 Human Experiment

To annotate the GELP dataset and probe humans' language processing, we conduct a crowdsourcing experiment. We explore to what extent plausibility types, construction types, and degrees of memory load lead to good-enough language processing.

[4] The complete list of verbs appears in Appendix A.
[5] https://platform.openai.com/docs/models
[6] We provide example prompts in Appendix B.

4.1 Methods

We collect three responses per item in GELP on Amazon Mechanical Turk.[7] We run our experiment on PCIbex.[8] Our results include data from 304 English native speakers.

Following previous psycholinguistic research (Christianson et al., 2001, 2006, 2017), our experiment uses a yes/no question-answering task instead of an NLI task, which our model evaluation uses. We select this task because its procedure is simple to understand and natural to probe humans' language understanding. In addition, the NLI task and yes/no question-answering task are highly interchangeable in the context of this study because to use GELP in the human experiment, all we have to do is to convert premises in GELP into polar questions (e.g., *Did the boy kick the ball?*), and to convert *entailment* and *non-entailment* labels into *yes* and *no* responses, respectively. Because of this high interchangeability, we believe that the task difference does not hinder the human–model comparison.

Figure 2 presents an example trial sequence in our experiment. First, a 1,000 ms fixation occurs to solicit workers' attention. Then, a context, which corresponds to a premise in GELP, appears in full. A worker presses the spacebar to indicate that s/he has finished reading the sentence. At this point, the sentence disappears. After a 500 ms interval, a yes/no question appears, which the worker answers by pressing J or F for *yes* or *no*, respectively. A worker repeats this procedure 96 times.[9]

To calculate accuracy, we first assign each item a *human answer* that represents the majority of the three responses and then determine whether it is equal to the pre-defined *correct answer*. The assignment of the human answer allows us to ensure that if an assigned human answer does not match a correct answer, the difference comes from misinterpretation rather than accident.

Given previous psycholinguistic research reviewed in Section 2.1, we expect that humans exhibit low accuracy due to good-enough sentence processing when they (i) process implausible descriptions, (ii) face a memory demand, and (iii) process certain implausible constructions relative to others.

[7] https://www.mturk.com

[8] https://farm.pcibex.net

[9] Appendix C reports more details about the experimental methods.

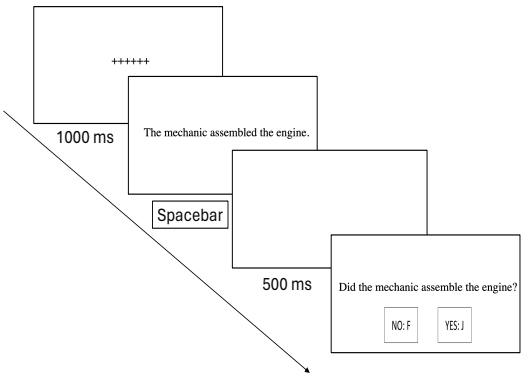


Figure 2: An example trial sequence in the human experiment.

4.2 Results

The gray bars in Figure 3 represent accuracy for humans. The average accuracy is 86.6%. This indicates that GELP is moderately challenging (cf. accuracy = 92 and 76% on MNLI and Heuristic Analysis for NLI Systems (HANS), respectively, (Nangia and Bowman, 2019; McCoy et al., 2019)).

Items with the correct *yes* answer have higher accuracy than those with the correct *no* answer (96.7 vs. 76.5%). We reason that humans may build a shallow syntactic representation of contexts; consequently, they tend to select *yes* when the context and question exhibit word overlap.

We saw little difference between the items with plausible contexts and those with implausible ones (87.9 vs 85.4%), contrary to our expectation (i). This indicates that humans do not adopt canonical form–meaning mapping or plausibility heuristics presumably because they adopt only the shallow syntactic analysis throughout our experiment.

The accuracy decreases as the memory load increases (92.8, 86.2, and 80.9% for low, medium, and high memory load conditions, respectively), confirming our expectation (ii). This result suggests that shallow language processing comes from a task-related memory demand. That is, humans adopt a shallow processing strategy to save cognitive resources so that they can memorize multiple propositions for subsequent question answering.

A closer look reveals that the accuracy is at the ceiling on the items with the correct *yes* answer (min. = 94.5%) but decreases on those with the correct *no* answer. The decrease depends on the memory load (86.9, 75.4, and 67.3% for low, medium, and high loads, respectively). This result should not be attributable to fallible memory. This is because the accuracy on the items with the correct *yes*

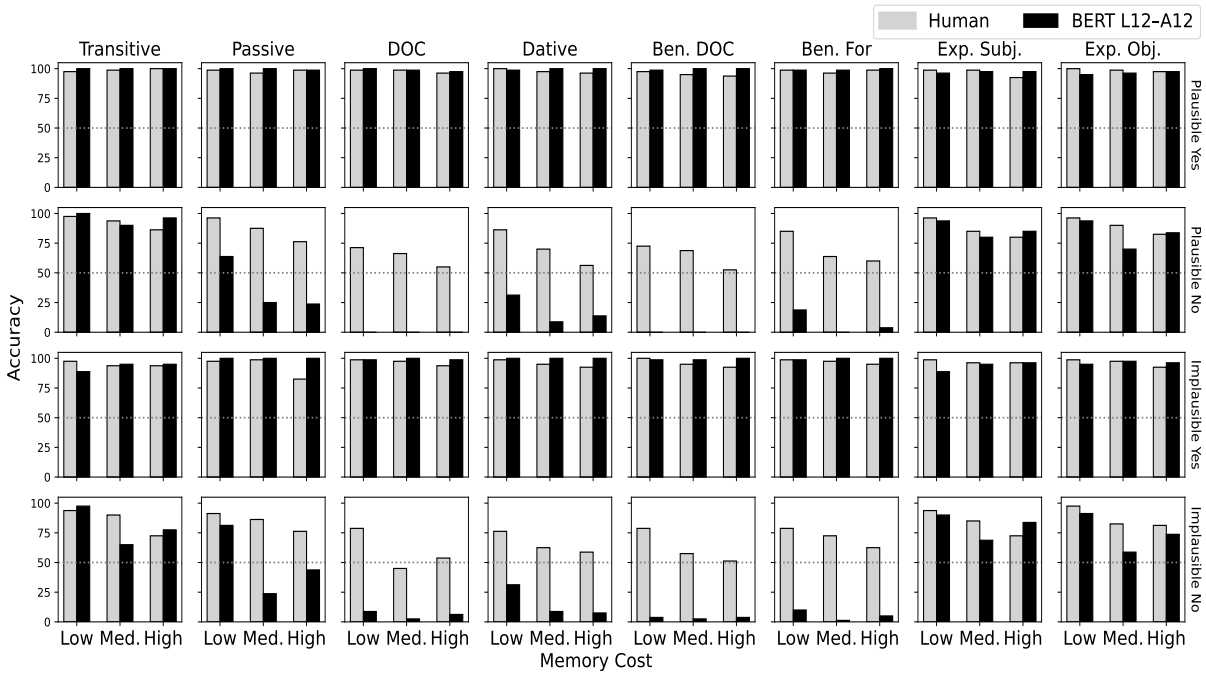


Figure 3: Accuracy for humans and BERT L12–A12 on GELP. The dotted lines indicate chance-level performance (50%).

answer is high regardless of the memory cost, indicating that humans can encode and store linguistic input in memory. A more viable explanation of the result would be that they do not engage in a detailed syntactic analysis in the first place under a memory-demanding situation; as a result, they encode words but not a detailed syntactic representation in working memory.

Finally, the accuracy on each construction varies but the variation is not specific to implausible sentences, contrary to our expectation (iii). Specifically, the accuracy on target sentences with two nouns is higher than that on those with three nouns (91.9 vs. 81.6% on average for transitive, passive, experiencer subject, and experiencer object constructions, on the one hand, and double object, dative, benefactive double object, and benefactive *for* constructions, on the other) regardless of plausibility (93.5 vs. 82.3% and 90.2 vs. 80.5% for plausible and implausible contexts of each group, respectively). Unlike Gibson et al. (2013), we use items with more than one proposition and present the context and question separately. We reason that the memory demand due to this task design facilitates a shallow syntactic analysis, which overrides another type of language processing observed in Gibson et al. (2013). The similarity in the accuracy pattern between the plausible and implausible conditions is not surprising if humans adopt the shallow syntactic analysis rather than other strategies throughout the experiment.

In summary, we confirm our expectation (ii) that the accuracy drops as the memory cost increases, suggesting that humans adopt good-enough sentence processing in a memory-demanding situation. However, we do not confirm the other two expectations (i) that implausible items have lower accuracy than plausible ones and (iii) that this difference depends on constructions.

5 Model Evaluation

Against the annotated GELP, we evaluate models with different architectural features. Assuming that a good-enough model neither engages in a detailed syntactic analysis nor needs a strong working memory, we hypothesize that our aimed model has (i) a shallow architecture and (ii) a small number of self-attention heads relative to the full model.

5.1 Models

We use Huggingface's (Wolf et al., 2020) 24 BERT miniatures (Turc et al., 2019), which cross six numbers of layers (L $\in \{2, 4, 6, 8, 10, 12\}$) with four numbers of self-attention heads (A $\in \{2, 4, 8, 12\}$). Turc et al. (2019) set four hidden embedding sizes (H $\in \{128, 256, 512, 768\}$) for each number of heads. We hereafter denote models as BERT Ln–An (e.g., BERT L12–A12).

Our evaluation uses an NLI task instead of the question-answering task because more training data are available for the former than the latter. As noted in Section 4.1, the task difference should not hinder us from evaluating models' good-enough language processing. We finetune the 24 BERT models on two standard NLI training datasets—SNLI (Bowman et al., 2015) and MNLI (Williams et al., 2018)—and a training split of HANS (McCoy et al., 2019). The inclusion of HANS intends to prevent the models from adopting structural heuristics all the time. Without HANS, we find that the models perform at chance on GELP by predicting *entailment* most of the time due to lexical overlap between a premise and hypothesis (e.g., 53% accuracy for BERT-base fine-tuned without HANS).

5.2 Evaluation Metric

Our evaluation metric is what we call a *human–model matching score*. This is calculated based on how many predicted labels match human responses. *Entailment* and *non-entailment* labels in the model evaluation correspond to *yes* and *no* responses in the human experiment, respectively. This compatibility between NLI labels and question-answering responses allows us to directly compare human and model results. Importantly, the human–model matching score is not the same as accuracy. For instance, the human–model matching score is 0.0 while the model accuracy is 1 if the model correctly labels the item while humans respond incorrectly.

5.3 Results

Table 3 presents the calculated human–model matching scores for each model. Although not perfect, the best-performed model is BERT L12–A12 (matching score = 74.3%), which corresponds to BERT-base. Six models with fewer layers and/or heads (BERTs L{4, 6, 8, 12}–A{8, 12} excluding BERTs L4–A8, L6–A8 and L6–A12) show 70% or higher scores (range: 70.0–73.2%). The comparable performance among these models suggests that good-enough language models do not require large architectures, confirming our overall hypothesis.

On average, the matching score improves from 54.9 to 68.7% and from 58.7 to 64.7% as layers and heads increase, respectively. Importantly, the difference among Ls = 8, 10, 12 is small (63.7, 65.3, 64.7%, respectively). This suggests that a deep architecture is not necessary for a good-enough model, which is consistent with our expectation (i). To explore this consideration, we finetune BERT L24–H16, which corresponds to BERT-large, in the same way as the 24 BERT models. We find that the matching score and accuracy for this model are 76.7 and 73.4%, respectively. The 2.3 vs. 4.2% improvement from BERT-base to BERT-large in these matrices suggests that a deep architecture contributes to the accuracy but less so to the human-like good-enough performance. Although a more in-depth analysis is necessary, this finding seems to be consistent with our interpretation of the model depth.

To analyze the effect of the number of attention heads, we calculate the model accuracy on GELP based on the memory cost (Table 4). We only report the results for models with 8 or 12 heads because most models with fewer heads perform at around the chance level.[10] The result shows that models exhibit decreasing accuracy as the memory cost increases. However, this accuracy pattern does not depend on the number of attention heads. For instance, the accuracy for the full BERT L12–A12 model was 74.1, 68.3, and 65.1% on low, medium, and high memory load conditions, respectively. This suggests that the number of attention heads does not necessarily contribute to models' good-enough performance, which is inconsistent with our hypothesis (ii). To further explore this conjecture, we fine-tune BERT-large, which corresponds to BERT L24–A16, in the same way as 24 BERT models. The fine-tuned BERT-large shows a similar pattern as the BERT L12–A12: the accuracy decreases as the memory load increases (77.9, 77.1, and 75.0% on low, medium, and high memory load conditions, respectively) but the decrease is more moderate than that observed in BERT-base. Since BERT-large differs from BERT-base in terms of hyperparameters other than the number of heads such as the model depth, we leave open what leads to this moderate decrease.

BERT L12–A12 shows the best matching score, 74.3%, and we find that its accuracy based on the pre-defined correct label is 69.2%, which is below human accuracy 86.6%. To disentangle what makes the model similar or different from the humans, Figure 3 presents the accuracy for the BERT L12–A12 on each condition in the black bars. We find that it performs at the ceiling on items with correct *entailment* (= *yes*) labels in a similar way as humans. However, its performance on items with correct *non-entailment* (= *no*) depends on construction

[10] Appendix D presents the full results.

L/A	2	4	8	12	Avg.
2	59.1 (0.6)	58.6 (0.6)	59.0 (0.6)	57.9 (0.6)	58.7 (0.6)
4	57.8 (0.6)	55.7 (0.6)	60.0 (0.6)	70.4 (0.5)	59. (0.6)
6	54.3 (0.6)	52.6 (0.6)	65.3 (0.5)	67.8 (0.5)	60.0 (0.6)
8	54.3 (0.6)	58.9 (0.6)	70.0 (0.5)	71.4 (0.5)	63.7 (0.6)
10	54.0 (0.6)	65.6 (0.5)	71.0 (0.5)	71.0 (0.5)	65.3 (0.5)
12	49.9 (0.6)	61.4 (0.6)	73.2 (0.6)	74.3 (0.5)	64.7 (0.6)
Avg.	54.9 (0.6)	58.8 (0.6)	66.4 (0.6)	68.8 (0.5)	

Table 3: Human–model matching score for 24 BERT models with different numbers of layers (L) and self-attention heads (A). Standard errors are in the parentheses.

L/A	8			12		
Memory load	Low	Medium	High	Low	Medium	High
2	50.9 (2.8)	50.6 (2.8)	50.2 (2.8)	49.1 (2.8)	50.5 (2.8)	49.8 (2.8)
4	54.7 (2.8)	51.9 (2.8)	52.9 (2.8)	69.4 (2.6)	64.8 (2.7)	63.6 (2.7)
6	59.0 (2.8)	58.0 (2.8)	56.6 (2.8)	62.7 (2.7)	61.8 (2.7)	61.4 (2.7)
8	66.1 (2.7)	64.8 (2.7)	61.3 (2.7)	69.2 (2.6)	66.3 (2.7)	64.8 (2.7)
10	66.3 (2.6)	65.6 (2.7)	64.6 (2.7)	68.5 (2.6)	64.6 (2.7)	62.9 (2.7)
12	69.3 (2.6)	68.1 (2.6)	67.7 (2.6)	74.1 (2.5)	68.3 (2.6)	65.1 (2.7)
Avg.	61.1 (2.7)	59.8 (2.7)	58.9 (2.7)	65.5 (2.6)	62.7 (2.7)	61.3 (2.7)

Table 4: Accuracy for BERT with 8 or 12 heads on GELP based on the three degrees of memory cost. Standard errors are in the parentheses.

types. Specifically, BERT L12–A12 performs well on the constructions with two-place predicates (i.e., transitive, experiencer subject, and experiencer object constructions; 83.3%) but at around chance on passives (43.5%) and poorly on constructions with three-place predicates (double object, dative, benefactive *for*, and benefactive double object constructions; 7.0%). The model's performance in the latter two does not resemble humans' although they are also not perfect (85.6 and 66.0% for passives and three-place predicate constructions, respectively).

The models' poor performance on passives and three-place predicates has a broad implication. Recent studies suggest that language models can learn to be sensitive to word order (Papadimitriou et al., 2022; Kauf et al., 2023). However, our results indicate that models' word order sensitivity can be specific to active sentences with two-place predicates. To explore whether the observed poor performance comes from the model's internal architecture or lack of relevant examples in training data, we retrain BERT L12–H12 on data augmented with 800 examples similar to passives or ditransitive constructions in GELP. The retrained model achieves the 84.3% human–model matching score. Although we cannot draw a strong conclusion because the training data that we use resembles the items in GELP, this result suggests that use of appropriate training data leads models to learn robust syntactic generalizations (McCoy et al., 2019).

In summary, the full model as well as models with fewer layers and/or heads show good-enough performance. The smaller number of layers does not considerably impair the models' good-enough language processing, confirming our first hypothesis (i) that a shallow architecture leads to a shallow representation of linguistic input. In contrast, the contribution of the number of heads is unclear, which does not confirm our second hypothesis (ii) that fewer heads parallel humans' limited working memory system.

6 Discussion

Does a shallow architecture lead to good-enough language processing? Our first hypothesis is that a good-enough model requires a shallow architecture because it does not have to make a detailed syntactic representation of linguistic input. We

confirm this hypothesis by finding that the shallow models exhibit a human-like good-enough performance in a way similar to their deeper version. Specifically, increasing the number of layers from 8 to 12 does not improve the models' human-like performance considerably.

Our results shed light on whether language models can learn the dissociation between formal linguistic competence—knowledge of grammar—and functional linguistic competence—the ability to use language in real-world situations (Mahowald et al., 2023). The advent of seemingly well-behaved neural language models leads to an intensive investigation of their formal linguistic competence (Marvin and Linzen, 2018; Futrell et al., 2019; Warstadt et al., 2019b,a, 2020). Mahowald et al. (2023) conclude that language models show promising results in learning abstract linguistic rules and patterns, but it remains to be seen whether they can learn functional linguistic competence. Our results point to the possibility that deep as well as shallow models can learn human-like language processing that is good-enough for simple language use.

Do fewer heads lead to good-enough language processing? Our second hypothesis is that the aimed model requires fewer attention heads because it does not need a strong working memory system. We find that among our model set, the model with the largest number of heads (H = 12) shows decreasing accuracy as the memory load increases, in a similar way as humans. This result does not confirm our hypothesis.

We can explain this result if the parallelism between the self-attention mechanism in Transformers and the human working memory system is specific to the retrieval phase. The human working memory system involves three phases, encoding, storage, and retrieval, at a coarse level of granularity (Baddeley, 1986, 2000). The models as well as humans achieve accuracy at the ceiling on items with correct *entailment/yes* labels, suggesting that they can store words that appear in contexts/premises. Thus, the observed accuracy pattern might not have to do with the storage phase of the working memory. We then conjecture that good-enough language processing has to do with the encoding phase of the working memory: the detailed syntactic analysis does not take place in the first place due to a high memory load during this phase, and as a result, the working memory does not store the detailed syntactic representation. We leave open the exact mechanism of this process and its connection to the self-attention architecture in Transformers for future research.

How does this study inform psycholinguistics? Although it is hard to make a direct comparison between humans and language models, it is worth considering if our findings can inform psycholinguistics in a meaningful way. Prior psycholinguistic studies postulate multiple possible sources of humans' good-enough performance. Some representative examples are semantic or structural heuristics (Ferreira, 2003), a shallow representation of linguistic input (Sanford and Sturt, 2002; Sanford and Graesser, 2006), and working memory burden (Christianson et al., 2006; Patson et al., 2006). The results from our model evaluation against human data are at least consistent with the view that humans adopt a non-detailed syntactic analysis. As we stipulated in the preceding paragraph, they also point to the possibility that it is the encoding but not storage or retrieval phase of the working memory system that is related to a source of humans' shallow sentence processing. Combining these two considerations suggests that the working memory demand during the encoding phrase leads to the shallow syntactic analysis of sentences (cf. Christianson et al., 2006).

7 Conclusion

From a cognitive perspective, we take language models' heuristic performance as their potential to learn human-like good-enough performance in language processing. This study creates a good-enough language processing evaluation dataset, GELP. We explore what architectural features contribute to models' good-enough language processing, focusing on the numbers of layers and self-attention heads. The model evaluation reveals that models with fewer layers and/or heads exhibit a good-enough performance in a similar way as a full model. This result leads us to conclude that the shallow architecture makes the models engage in an underspecified syntactic analysis. In contrast, the role of the self-attention mechanism is unclear. We leave open its exact role in the model's good-enough language processing. We hope that this study will foster more psycholinguistically oriented research on language models' good-enough performance.

Limitations

This study has three limitations. First, our model coverage is limited. We evaluate only BERT with different architectural features. Therefore, it remains to be seen whether our findings and hypotheses are generalizable to other Transformer-based models such as GPT-2.

Second, the use of templates in the dataset creation may result in unnatural sentences in an unintended way such that it leads us to fail to measure what we intend to measure (e.g., the effect of the construction types).

Finally, this study does not test whether models can perform in a human-like way in terms of both grammatical knowledge and good-enough language processing. According to Ferreira (2003), humans have a good balance between robust, algorithmic language processing and non-robust, heuristic language processing so that they do not always misinterpret linguistic input. It is necessary to test whether models can learn the human-like balance.

Ethics Statement

When we use Amazon Mechanical Turk, we make sure that our payment and rejection policies are reasonable and comparable to in-person employment. The task that we use in our crowdsourcing experiment is a simple yes/no question-answering task; hence, it should cause no harm to workers. This work passes review from the oversight of the internal review boards of the authors' institutes.

Acknowledgements

We wish to thank the anonymous reviewers for their constructive feedback. This work was supported by JST PRESTO Grant Number JPMJPR20C4 and JSPS KAKENHI Grant Number 22K17954.

References

Alan D. Baddeley. 1986. *Working memory*. Oxford University Press, Oxford: UK.

Alan D. Baddeley. 2000. The episodic buffer: A new component of working memory? *Trends in cognitive sciences*, 4(11):417–423.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. arXiv preprint 2005.14165.

Kiel Christianson. 2016. When language comprehension goes wrong for the right reasons: Good-enough, underspecified, or shallow language processing. *Quarterly journal of experimental psychology*, 69(5):817–828.

Kiel Christianson, Andrew Hollingworth, John F Halliwell, and Fernanda Ferreira. 2001. Thematic roles assigned along the garden path linger. *Cognitive psychology*, 42(4):368–407.

Kiel Christianson, Steven G Luke, and Fernanda Ferreira. 2010. Effects of plausibility on structural priming. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 36(2):538.

Kiel Christianson, Steven G Luke, Erika K Hussey, and Kacey L Wochna. 2017. Why reread? evidence from garden-path and local coherence structures. *Quarterly Journal of Experimental Psychology*, 70(7):1380–1405.

Kiel Christianson, Carrick C Williams, Rose T Zacks, and Fernanda Ferreira. 2006. Younger and older adults'" good-enough" interpretations of garden-path sentences. *Discourse processes*, 42(2):205–238.

Cleo Condoravdi, Dick Crouch, Valeria de Paiva, Reinhard Stolle, and Daniel G. Bobrow. 2003. Entailment, intensionality and text understanding. In *Proceedings of the HLT-NAACL 2003 Workshop on Text Meaning*, pages 38–45.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Mengnan Du, Fengxiang He, Na Zou, Dacheng Tao, and Xia Hu. 2023. Shortcut learning of large language models in natural language understanding.

Evelina Fedorenko, Idan Asher Blank, Matthew Siegelman, and Zachary Mineroff. 2020. Lack of selectivity for syntax relative to word meanings throughout the language network. *Cognition*, 203:104348.

Fernanda Ferreira. 2003. The misinterpretation of noncanonical sentences. *Cognitive psychology*, 47(2):164–203.

Fernanda Ferreira and Nikole D. Patson. 2007. The 'good enough' approach to language comprehension. *Language and linguistics compass*, 1(1-2):71–83.

Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 2019. Neural language models as psycholinguistic subjects: Representations of syntactic state. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 32–42. Association for Computational Linguistics.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

Edward Gibson, Leon Bergen, and Steven T Piantadosi. 2013. Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences*, 110(20):8051–8056.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112. Association for Computational Linguistics.

Thilo Hagendorff and Sarah Fabi. 2023. Why we need biased ai: How including cognitive biases can enhance ai systems. *Journal of Experimental & Theoretical Artificial Intelligence*, pages 1–14.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031. Association for Computational Linguistics.

Carina Kauf, Anna A Ivanova, Giulia Rambelli, Emmanuele Chersoni, Jingyuan Selena She, Zawad Chowdhury, Evelina Fedorenko, and Alessandro Lenci. 2023. Event knowledge in large language models: the gap between the impossible and the unlikely. *Cognitive Science*, 47(11):e13386.

Yuxuan Lai, Chen Zhang, Yansong Feng, Quzhe Huang, and Dongyan Zhao. 2021. Why machine reading comprehension models learn shortcuts? In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 989–1002. Association for Computational Linguistics.

Beth Levin. 1993. *English verb classes and alternations: A preliminary investigation*. University of Chicago press.

Tal Linzen. 2020. How can we accelerate progress towards human-like linguistic generalization? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5210–5217. Association for Computational Linguistics.

Kyle Mahowald, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2023. Dissociating language and thought in large language models: a cognitive perspective.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202. Association for Computational Linguistics.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448. Association for Computational Linguistics.

Aaron Mueller and Tal Linzen. 2023. How to plant trees in language models: Data and architectural effects on the emergence of syntactic inductive biases. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11237–11252. Association for Computational Linguistics.

Nikita Nangia and Samuel R. Bowman. 2019. Human vs. muppet: A conservative estimate of human performance on the GLUE benchmark. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4566–4575. Association for Computational Linguistics.

Isabel Papadimitriou, Richard Futrell, and Kyle Mahowald. 2022. When classifying arguments, BERT doesn't care about word order...except when it matters. In *Proceedings of the Society for Computation in Linguistics 2022*, pages 203–205. Association for Computational Linguistics.

Nikole D. Patson, E Swensen, N Moon, and Fernanda Ferreira. 2006. Individual differences in syntactic reanalysis. Poster presented at the Annual CUNY Conference on Human Sentence Processing.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Soo Hyun Ryu and Richard Lewis. 2021. Accounting for agreement phenomena in sentence comprehension with transformer language models: Effects of similarity-based interference on surprisal and attention. In *Proceedings of the Workshop on Cognitive*

Modeling and Computational Linguistics, pages 61–71. Association for Computational Linguistics.

Anthony J Sanford and Arthur C Graesser. 2006. Shallow processing and underspecification. *Discourse Processes*, 42(2):99–108.

Anthony J Sanford and Patrick Sturt. 2002. Depth of processing in language comprehension: Not noticing the evidence. *Trends in cognitive sciences*, 6(9):382–386.

Saku Sugawara, Kentaro Inui, Satoshi Sekine, and Akiko Aizawa. 2018. What makes reading comprehension questions easier? In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4208–4219. Association for Computational Linguistics.

Ruixiang Tang, Dehan Kong, Longtao Huang, and Hui Xue. 2023. Large language models can be lazy learners: Analyze shortcuts in in-context learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4645–4657. Association for Computational Linguistics.

William Timkey and Tal Linzen. 2023. A language model with limited memory capacity captures interference in human sentence processing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8705–8720. Association for Computational Linguistics.

Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Well-read students learn better: On the importance of pre-training compact models. *arXiv preprint arXiv:1908.08962v2*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Alex Warstadt, Yu Cao, Ioana Grosu, Wei Peng, Hagen Blix, Yining Nie, Anna Alsop, Shikha Bordia, Haokun Liu, Alicia Parrish, Sheng-Fu Wang, Jason Phang, Anhad Mohananey, Phu Mon Htut, Paloma Jeretic, and Samuel R. Bowman. 2019a. Investigating BERT's knowledge of language: Five analysis methods with NPIs. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2877–2887. Association for Computational Linguistics.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019b. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

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 the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20. Association for Computational Linguistics.

A Verb Lists

We use the following verbs for creating our dataset.

Verbs for transitives/passives *assemble, bend, bite, blend, carve, chop, clean, collect, cut, describe, design, destroy, draw, flatten, fold, hack, hammer, hit, kick, knock, make, memorize, pound, produce, protect, punch, push, read, saw, save, shatter, skip, slash, slice, smash, squash, suggest, touch, use, waste*

Verbs for DOCs/datives *allocate, assign, award, bring, email, extend, fax, feed, forward, give, grant, hand, haul, issue, lend, lease, leave, loan, mail, offer, owe, pass, pay, post, promise, refund, relay, repay, sell, send, serve, ship, show, slip, smuggle, take, teach, tell, trade, write*

Verbs for benefactive DOCs/benefactive *for* constructions *arrange, assemble, bake, book, boil, build, buy, carve, cash, catch, charter, clean, compile, cook, cut, design, develop, earn, find, fix, get, grill, grow, keep, knit, make, order, paint, pick, prepare, rent, reserve, roll, save, secure, set, shape, steal, wash, write*

Verbs for experiencer subject constructions *abhor, admire, adore, appreciate, cherish, covet, crave, deplore, desire, despise, disdain, dislike, distrust, dread, enjoy, envy, exalt, execrate, favor, fear, hate, lament, like, loathe, love, miss, mourn, need, pity, regret, relish, resent, savor, tolerate, treasure, trust, value, venerate, want, worship*

Verbs for experiencer object constructions *agonize, amaze, amuse, anger, annoy, arouse, astonish, bore, bother, calm, captivate, comfort, confuse, convince, depress, devastate, disappoint, discourage, disgust, disturb, displease, embarrass, encourage, enlighten, excite, frighten, frustrate, impress, irritate, please, puzzle, sadden, satisfy, shock, surprise, terrify, threaten, thrill, upset, worry*

B Example Prompts

A template for a prompt: Can you make [CONSTRUCTION NAME] with the following verbs?

Please...
1. Use an inanimate entity in [POSITION OF INTEREST].
2. Use an animate entity in [POSITION OF INTEREST].
3. Use past tense for the verb.
4. Use no pronouns.
5. Use no adjectives.

[LIST OF VERBS]

An example prompt for the creation of transitive sentences: Can you make transitive constructions with the following verbs?

Please...
1. Use an inanimate entity in the subject.
2. Use an animate entity in the object.
3. Past tense for the verb.
4. Use no pronouns.
5. Use no adjectives.

agonize, amaze, amuse, anger, annoy, arouse, astonish, bore, bother, calm, captivate, comfort, confuse, convince, depress, devastate, disappoint, discourage, disgust, disturb, displease, embarrass, encourage, enlighten, excite, frighten, frustrate, impress, irritate, please, puzzle, sadden, satisfy, surprise, shock, terrify, threaten, thrill, upset, worry

C Details on Human Experiment

C.1 Participants

Using Amazon Mechanical Turk, we recruit workers with the requirements of having an approval rating of 99.0% or higher, having at least 5,000 approved tasks, and being located in the US, the UK, or Canada. We calculate the reward as $12 per hour. We first recruit 1,200 workers for the qualification task. The qualification task has 20 context–question pairs, which resemble items in GELP. We take more than 70% accuracy in the qualification task as a threshold for the invitation to the actual experiment. Based on this exclusion criterion, we invite 487 workers to participate in the actual experiment. Among them, 304 workers take part.

Our experiment collects no personal information. By accepting PCIbex's participation agreement, workers consent to the collection and use of non-personal data for research purposes.

C.2 List Design

In addition to critical 7,680 items in GELP, our experiment includes 7,680 distractor pairs to ensure a balanced and unbiased assessment of humans' sentence processing. They serve to mitigate potential response strategies (e.g., paying attention to only a target context) by masking the critical items. Contexts in the distractors consist of two (N = 2,560) or three (N = 5,120) propositions, where one proposition corresponds to a target context; questions ask about a thematic relation in filler contexts. Half of them have a correct *yes* response; the other half have a correct *no* response.

We divide the 15,360 context–question pairs (7,680 targets + 7,680 distractors) into 160 lists, each of which has 96 items. The number of items in each list roughly follows previous psycholinguistic research (Christianson et al., 2001, 2006). We make sure that no worker sees the same list more than one time.

C.3 Instruction

We show workers an instruction in Figure 4 before the experiment to familiarize them with the experimental procedure.

D Model Accuracy by Memory Load

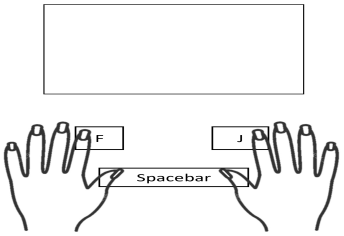
Table 5 presents full results on model accuracy by memory load.

Instruction

In this experiment, you will read a sentence and answer a yes/no question about its content.

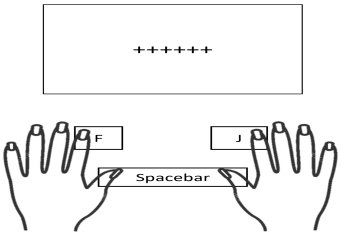
Finger position

You will use **F**, **J**, and **Spacebar**. During the experiment, please put your fingers as below.

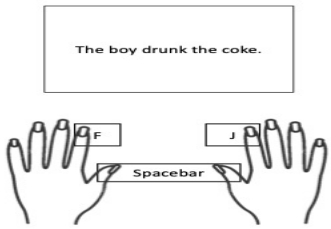


Task

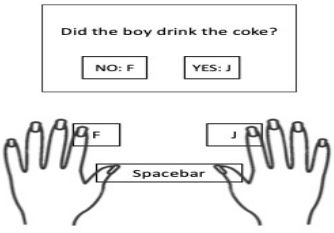
First, '+ + + + + +' appears on the screen.



Then, a sentence appears. Read it at a normal rate and press Spacebar when you finish reading it.



Finally, a yes/no question appears on the screen.



If your answewer to the question is yes, press J. If not, press F.

You will repeat this procedure **96** times.

Figure 4: Instruction for the experiment.

L/A	2			4			8			12		
Memory load	Low	Med.	High	Low	Med.	High	Low	Med.	High	Low	Med.	High
2	50.0	50.0	50.0	50.0	50.3	50.2	50.9	50.6	50.2	49.1	50.5	49.8
4	48.9	50.1	50.5	47.0	50.4	50.4	54.7	51.9	52.9	69.4	64.8	63.6
6	49.8	50.5	49.7	48.0	49.2	50.2	59.0	58.0	56.6	62.7	61.8	61.4
8	49.6	50.0	50.7	53.6	51.7	51.8	66.1	64.8	61.3	69.2	66.3	64.8
10	50.3	51.8	49.9	62.5	57.7	54.3	66.3	65.6	64.6	68.5	64.6	62.9
12	49.9	50.5	50.0	55.1	51.4	51.3	69.3	68.1	67.7	74.1	68.3	65.1
Avg.	49.8	50.4	50.1	52.7	51.8	51.4	61.1	59.8	58.9	65.5	62.7	61.3

Table 5: Accuracy for all models on GELP based on the three degrees of memory load.