CMCL 2024 (Findings of ACL 2024)

*daiasami@udel.edu

What Makes Language Models Good-enough?

*Daiki Asami (University of Delaware) Saku Sugawara (National Institute of Informatics)

Paper

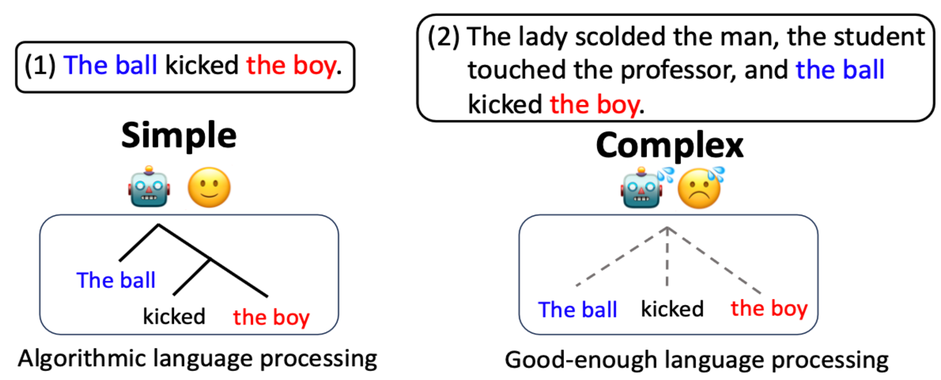
Background: Heuristic Language Processing

Language Models:

- LMs often adopt heuristics to make predictions.
  e.g., BERT predicts *entailment* for the following pair (McCoy+, 2019):
  Premise: *The doctor was paid by the actor.* Hypothesis: *The actor paid the doctor.*

Humans:

- Humans occasionally adopt *good-enough* heuristic language processing rather than detailed *algorithmic* language processing (Ferreira, 2003).
  e.g., They misinterpret *the dog was bitten by the teacher* as *the teacher was bitten by the dog*.
- Heuristic processing can occur under a cognitively demanding situation (e.g., sentence complexity).



Research Questions and Hypotheses

RQ1: To what extent is LMs' heuristic processing human-like?

RQ2: What architectural features contribute to their good-enough language processing?

Deep architecture enables LMs to learn a syntactic generalization (Mueller and Linzen, 2023). → The shallow architecture would lead to good-enough models.
Self-attention head mechanism resembles a working-memory system (Ryu and Lewis, 2021). → The small number of heads would contribute to good-enough models.

Good-enough Language Processing (GELP) Dataset for Natural Language Inference Task

7,680 items (80 sentences * 8 construction types * 3 working-memory loads * 2 premise plausibility types * 2 hypothesis types):

- 8 constructions: (1) transitive, (2) passive, (3) double object, (4) dative, (5) benefactive double object, (6) benefactive *for*, (7) experiencer subject, & (8) experiencer object constructions.
- We make 80 sentences for each construction.
- 3 memory loads: low, medium, and high memory load conditions with one, two, and three propositions in premises, respectively.

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.

- 2 plausibility types : implausibility caused by swapping animate and inanimate nouns.
- 2 hypothesis labels: *entailment* and *non-entailment* (covering *neutral* and *contradiction*).

Human annotation:

- We collect three human labels for each item on Amazon Mechanical Turk.
- We assign the *human label* to each item based on the majority label.

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)

Model Evaluation

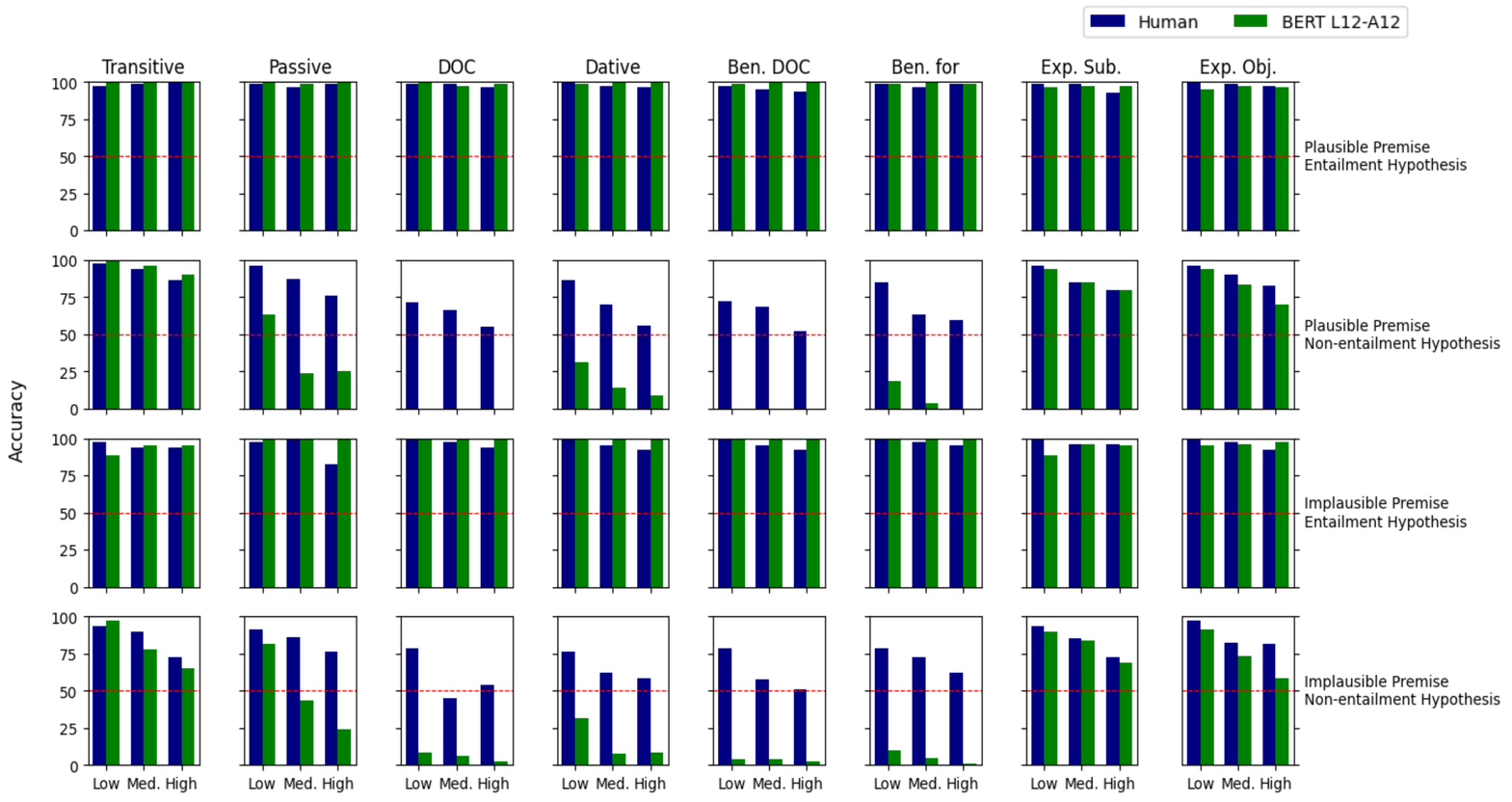
24 BERT miniatures (6 numbers of layers (L ∈ {2, 4, 6, 8, 10, 12}) * 4 numbers of self-attention heads (A ∈ {2, 4, 8, 12})) (Turc et al., 2019) fine-tuned on NLI task.

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)	

Human–model matching score (= the proportion of the match btw models' predicted labels and human labels) for 24 BERT models.

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)

Accuracy for BERT with 8 or 12 heads based on three memory loads.



Accuracy for humans and BERT L12–A12 (the best performing model).

Discussion

RQ1: To what extent is LMs' heuristic processing human-like?

- Taking 70% human–model matching score as a threshold, the full model as well as models with fewer layers and/or heads exhibit good-enough performance.

RQ2: What architectural features contribute to their good-enough language processing?

- Increasing layers (i.e., 8 to 12) does not improve the performance considerably. The shallow models exhibit a good-enough performance like their deeper version.
- The model with the largest number of heads (i.e., 12) shows decreasing accuracy as the memory load increases like humans. This result does not confirm our hypothesis that the aimed models require a fewer number of heads.

References

Ferreira. 2003. The Misinterpretation of Noncanonical Sentences. In Cognitive Psychology; McCoy et al. 2019. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. In ACL; Mueller and Linzen. 2023. How to Plant Trees in Language Models: Data and Architectural Effects on the Emergence of Syntactic Inductive Biases. In ACL; Ryu and Lewis. 2021. Accounting for Agreement Phenomena in Sentence Comprehension with Transformer Language Models: Effects of Similarity-based Interference on Surprisal and Attention in CMCL; Turc et al. 2021. Well-read Students Learn Better: On the Importance of Pre-training Compact Models. In arXiv.