Adjective Scale Probe: Can Language Models Encode Formal Semantics Information?

Wei Liu[1], Ming Xiang[2], Nai Ding[1]

[1]College of Biomedical Engineering and Instrument Sciences, Zhejiang University

[2]Department of Linguistics, The University of Chicago

浙江大学 ZHEJIANG UNIVERSITY

THE UNIVERSITY OF CHICAGO

Introduction

Background:

Current language models perform well on many tasks[1]

- Models surpassed human performance on the GLUE benchmark.

Models are susceptible to adversarial attacks[2].

- Model performance dropped by 30% by small perturbation.

Question:

Do such models truly understand the meaning of language or simply guessing answers?

[1] He P, Gao J, and Chen W. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing[J]. arXiv preprint, 2021.
[2] Wang B, Xu C, Wang S, et al. Adversarial glue: A multi-task benchmark for robustness evaluation of language models[C]. NeurIPS, 2022.

Introduction

Goal of the current study:

We formulate a theoretically motivated test for how well models can understand adjectives and adjective phrases.

The meaning of adjectives:

- Adjectives have highly context-sensitive meaning, which makes the meaning quite variable in different contexts.
- **Degree semantics** analysis[3][4] of adjectives postulates a semantic core underlying the meaning of all adjectives in all contexts.

[3] Cresswell M. The semantics of degree[J]. Montague Grammar, 1976.
[4] Kennedy C and McNally L. Scale Structure, Degree Modification, and the Semantics of Gradable Predicates[J]. Language, 2005.

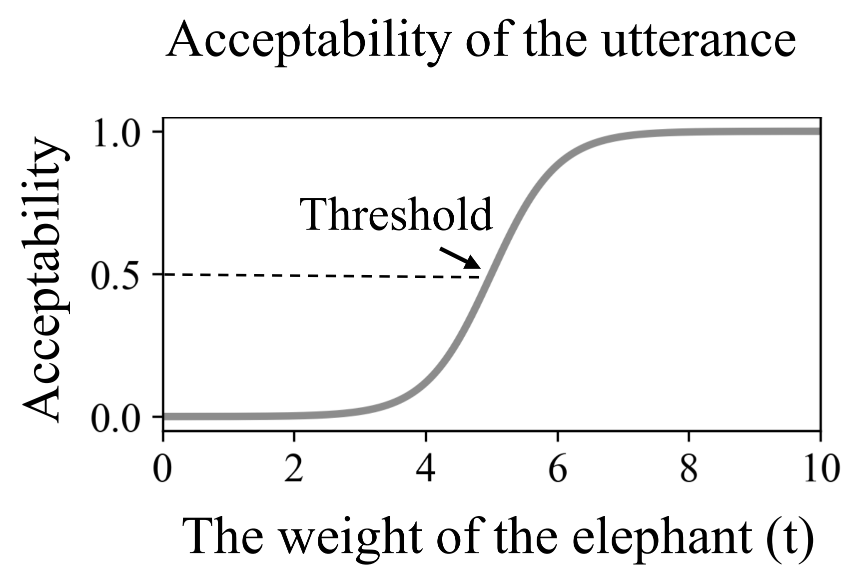
Introduction

Degree semantics:

*The elephant is **heavy**.*

Scale structure

- Argument (***elephant***)
- Dimension (***mass***)
- Context (***Threshold***)
- Ordering relation
- …



Can language models capture the degree semantics of adjectives?

Dataset construction

Adjective Scale Probe: test adjective interpretation using the natural language inference (NLI) task, based on the degree semantics.

Premise:

A 120 cm rod is long. A 80 cm rod is short.

Entail →

Hypothesis 1:

A 60 cm rod is short.

Not entail →

Hypothesis 2:

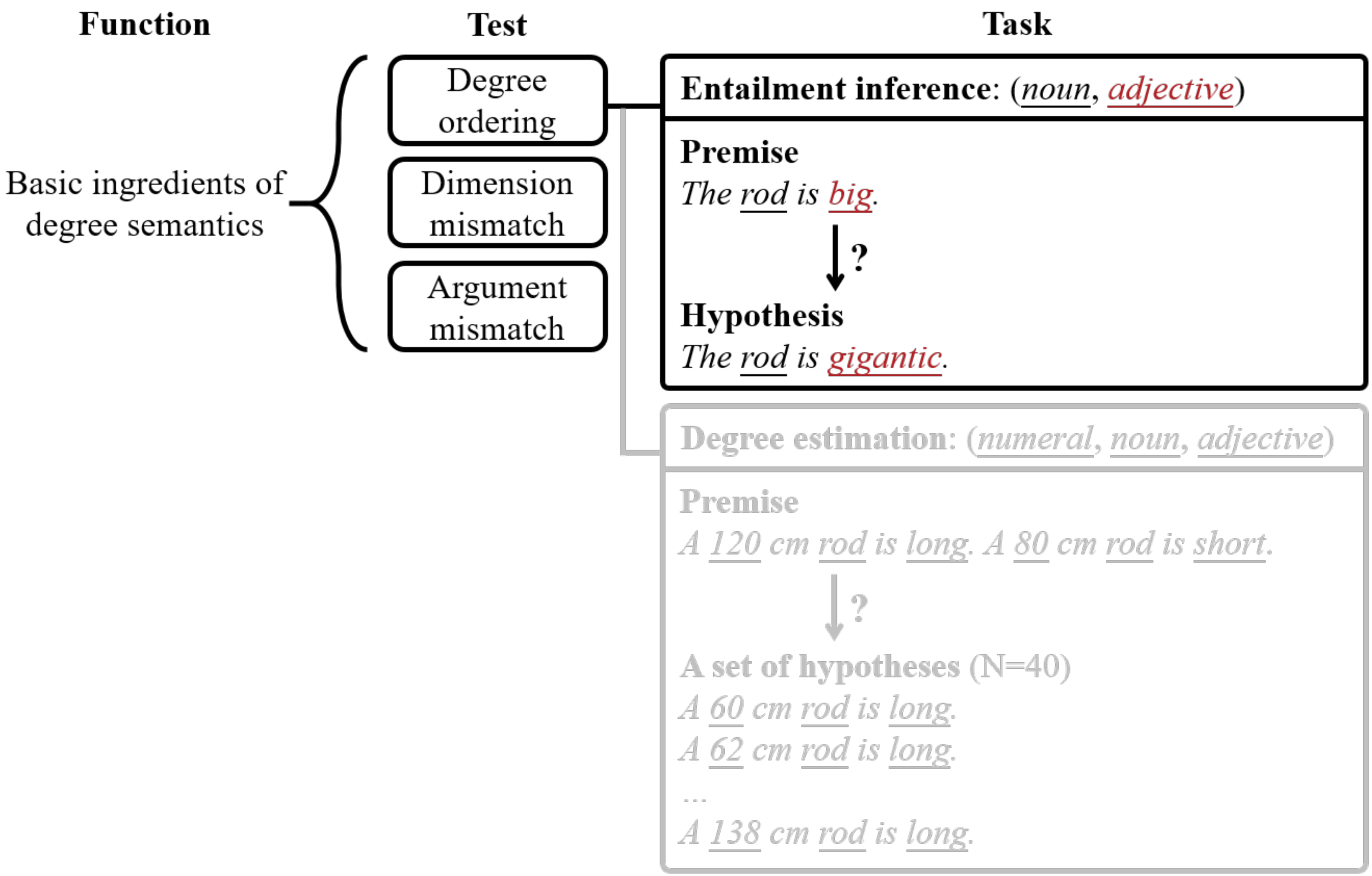
A 130 cm rod is short.

Dataset construction

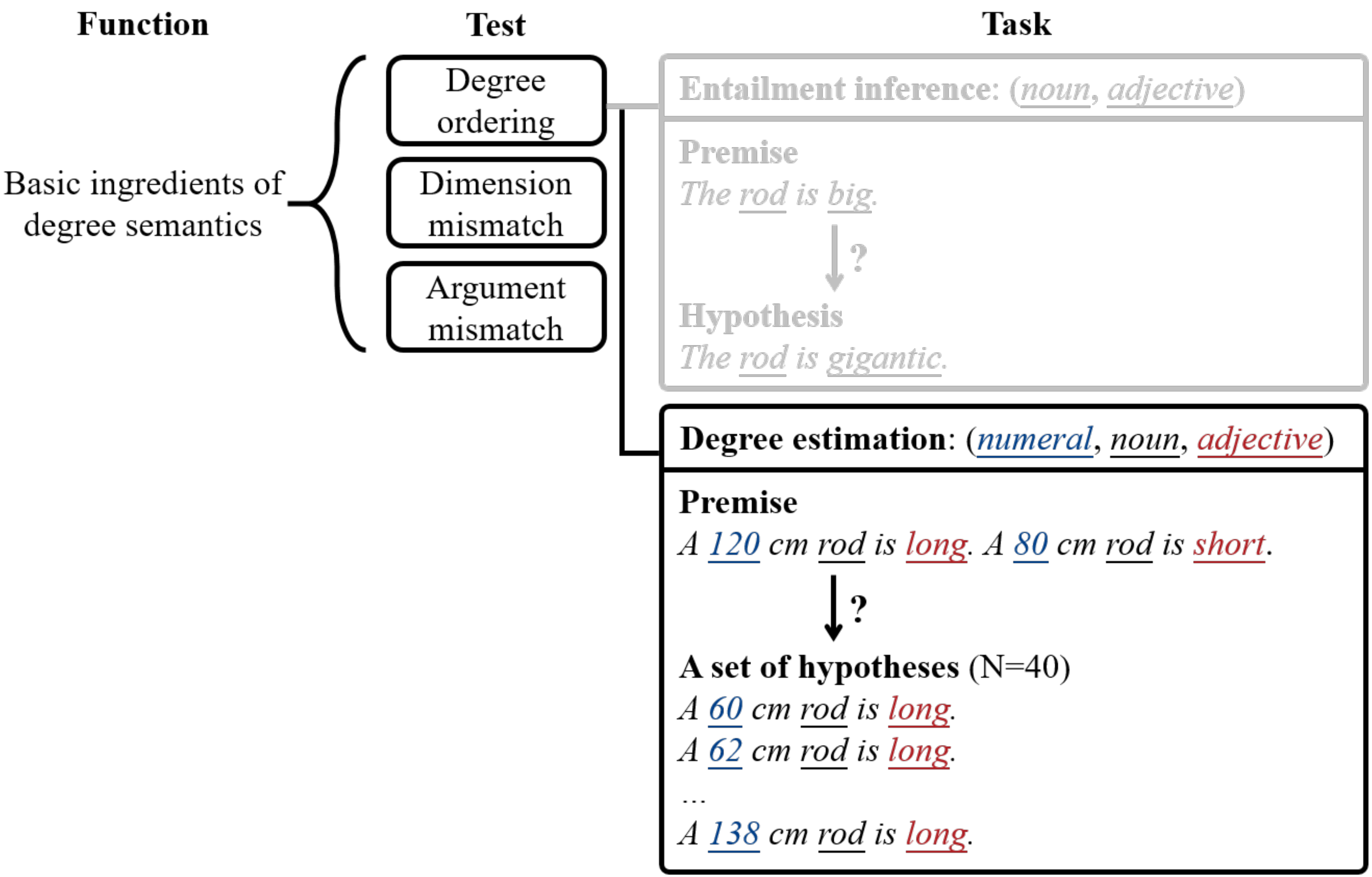
2 tasks, 8 tests for 3 aspects of adjective interpretation.

Function	Test	Task
Basic ingredients of degree semantics	Degree ordering	**Entailment inference**: (*noun*, *adjective*)
	Dimension mismatch	
	Argument mismatch	
Operations on degrees	Booster	**Degree estimation**: (*numeral*, *noun*, *adjective*)
	Diminisher	
	Negation	
Interpretation of degree morphemes (-er/more, -est/most)	Comparative	
	Superlative	

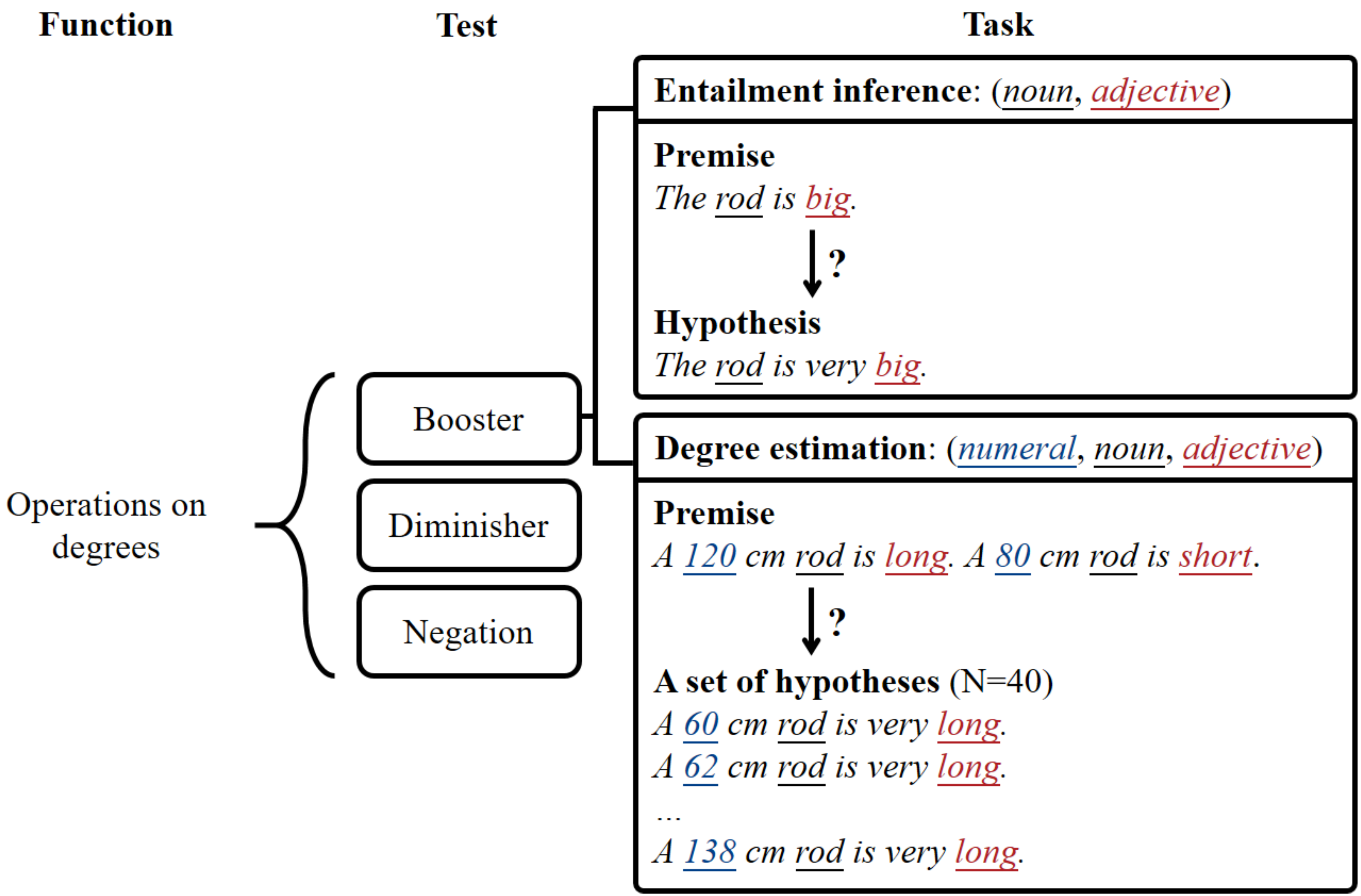
Dataset construction



Dataset construction



Dataset construction



Dataset construction

Function | **Test** | **Task**

Interpretation of degree morphemes (-er/more, -est/most)

- Comparative
- Superlative

Entailment inference: (*noun*, *adjective*)

Premise

The rod is bigger than the pole.

↓ ?

Hypothesis

The rod is big.

Degree estimation: (*numeral*, *noun*, *adjective*)

Premise

Rods longer than 120 cm is long. Rods shorter than 80 cm is short.

↓ ?

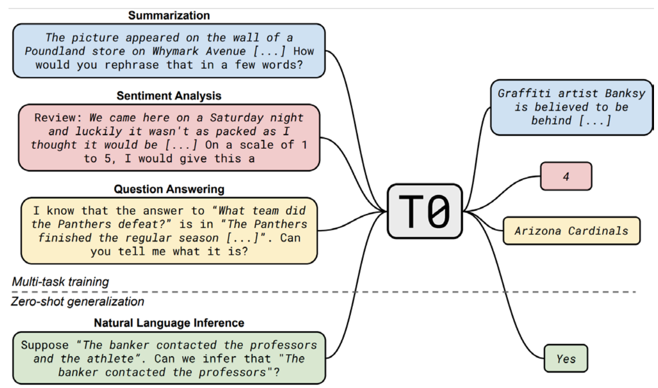
A set of hypotheses (N=40)

A 60 cm rod is long.

…

A 138 cm rod is long.

Experimental setup



BERT, DeBERTa **T0 (zero-shot)** **Human annotator**

Performance on entailment inference

Model	Entailment inference							
	Ingredient			**Operation**			**Morpheme**	
	Ord.	Dim.	Arg.	Bo.	Di.	Ne.	Com.	Sup.
BERT-base	56.9	89.9	60.5	52.7	30.8	70.6	44.7	44.2
BERT-large	53.7	87.7	67.0	51.4	32.7	**71.1**	43.8	43.1
DeBERTa-base	56.2	94.0	81.6	**59.8**	10.4	68.0	38.2	31.2
DeBERTa-large	**59.4**	96.1	85.6	55.8	3.8	67.9	47.6	55.2
T0 3B	52.2	**97.0**	85.2	50.7	48.3	43.4	**57.1**	50.6
T0 pp	57.2	94.6	**86.4**	50.1	**50.2**	64.7	55.6	**56.7**
Chance level	50.0	66.6	66.6	50.0	50.0	41.7	55.6	50.0
Majority baseline	50.0	100.0	100.0	50.0	50.0	75.0	66.7	50.0

The accuracy of the best performing model was never more than 10% above the majority baseline.

Performance on entailment inference

Model	Entailment inference							
	Ingredient			Operation			Morpheme	
	Ord.	Dim.	Arg.	Bo.	Di.	Ne.	Com.	Sup.
BERT-base	56.9	89.9	60.5	52.7	30.8	70.6	44.7	44.2
BERT-large	53.7	87.7	67.0	51.4	32.7	**71.1**	43.8	43.1
DeBERTa-base	56.2	94.0	81.6	**59.8**	10.4	68.0	38.2	31.2
DeBERTa-large	**59.4**	96.1	85.6	55.8	3.8	67.9	47.6	55.2
T0 3B	52.2	**97.0**	85.2	50.7	48.3	43.4	**57.1**	50.6
T0 pp	57.2	94.6	**86.4**	50.1	**50.2**	64.7	55.6	**56.7**
Chance level	50.0	66.6	66.6	50.0	50.0	41.7	55.6	50.0
Majority baseline	50.0	100.0	100.0	50.0	50.0	75.0	66.7	50.0

DeBERTa-large surpassed human performance on the NLU benchmarks[4], but performed poorly on the ASP.

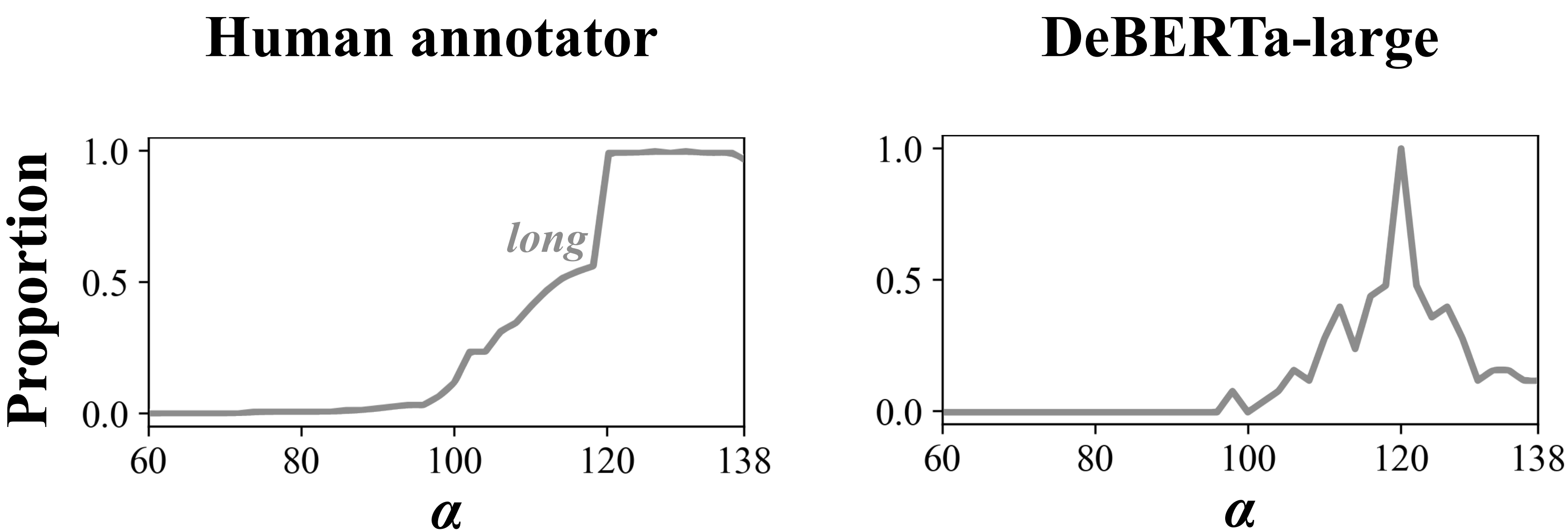
[4] He P, Gao J, and Chen W. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing[J]. arXiv preprint, 2021.

Performance on entailment inference

Model	Entailment inference							
	Ingredient			**Operation**			**Morpheme**	
	Ord.	Dim.	Arg.	Bo.	Di.	Ne.	Com.	Sup.
BERT-base	56.9	89.9	60.5	52.7	30.8	70.6	44.7	44.2
BERT-large	53.7	87.7	67.0	51.4	32.7	**71.1**	43.8	43.1
DeBERTa-base	56.2	94.0	81.6	**59.8**	10.4	68.0	38.2	31.2
DeBERTa-large	**59.4**	96.1	85.6	55.8	3.8	67.9	47.6	55.2
T0 3B	52.2	**97.0**	85.2	50.7	48.3	43.4	**57.1**	50.6
T0 pp	57.2	94.6	**86.4**	50.1	**50.2**	64.7	55.6	**56.7**
Chance level	50.0	66.6	66.6	50.0	50.0	41.7	55.6	50.0
Majority baseline	50.0	100.0	100.0	50.0	50.0	75.0	66.7	50.0

Language models failed to capture the degree difference between lexical items.

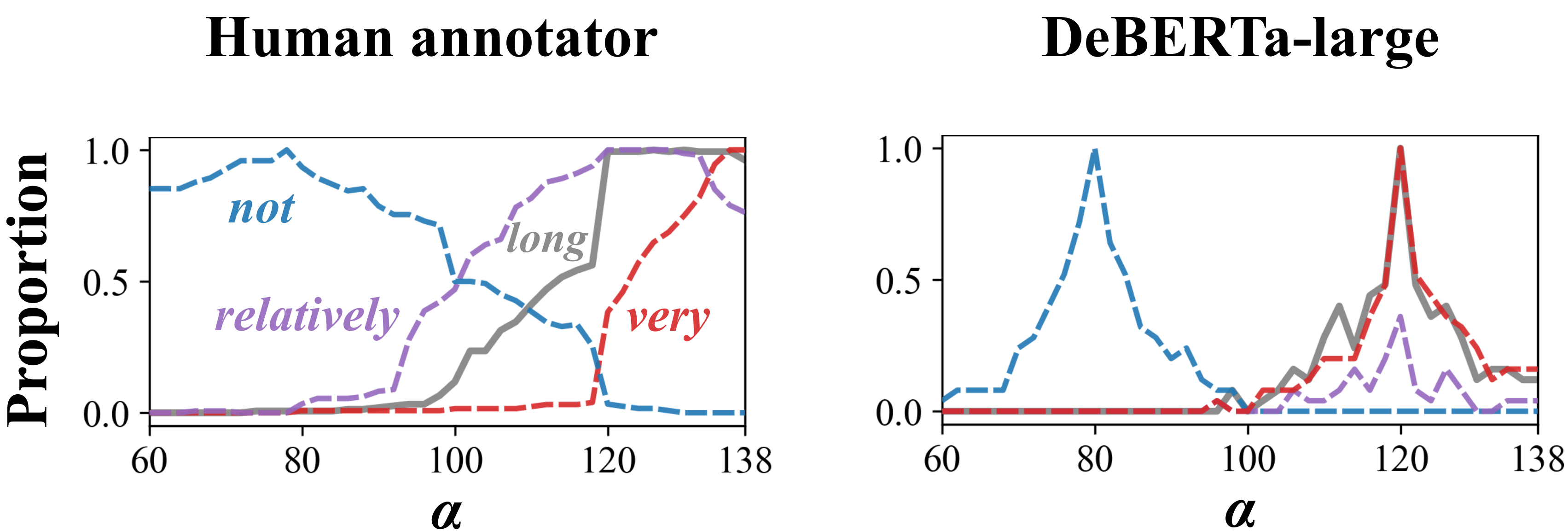
Performance on degree estimation



Premise: ***A 120 cm rod is long. A 80 cm rod is short.***

Hypotheses: ***A α cm rod is long.*** **α∈[60,138]**

Performance on degree estimation



Premise: ***A 120 cm rod is long. A 80 cm rod is short.***

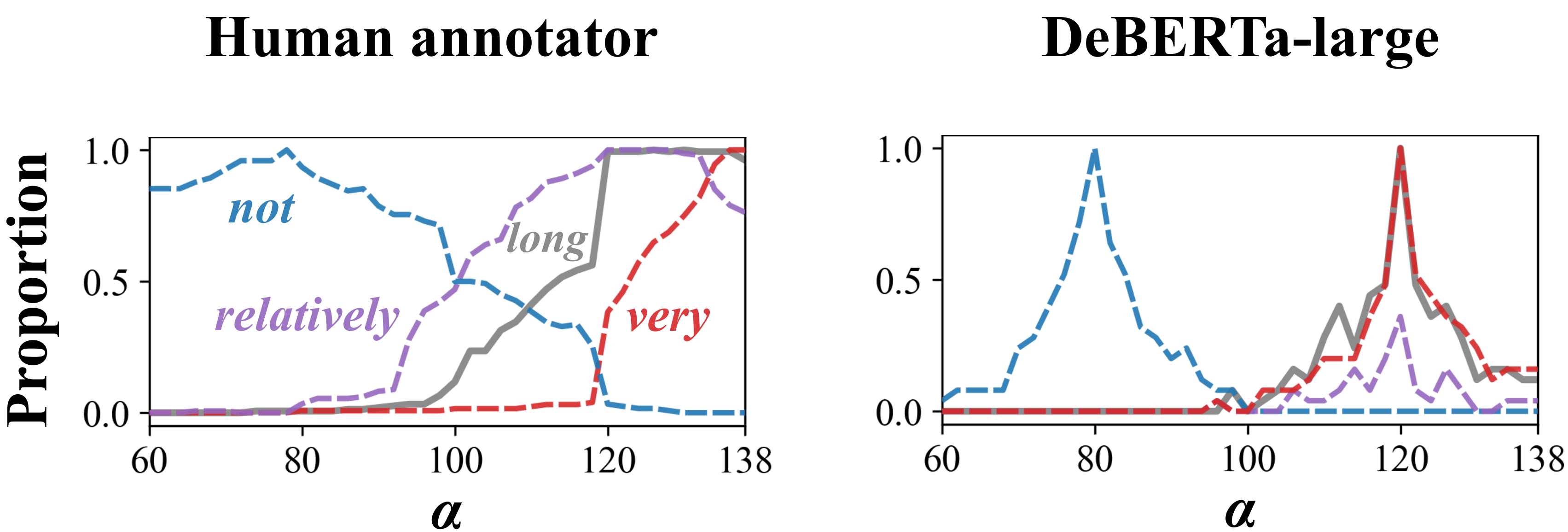
Hypotheses: ***A α cm rod is long.*** **$\alpha \in [60,138]$**

A α cm rod is very long.

A α cm rod is relatively long.

A α cm rod is not long.

Performance on degree estimation



Language models failed to understand the degree semantics of adjectives.

Fine-tuning on the ASP

Two possibilities for the poor performance of model:

- a) Language models do not encode the degree semantics.
- b) Language models encode the degree semantics, but fail to apply it in the current task.

We fine-tuned models on a subset of ASP.

- Models encode the degree semantics if the fine-tuning effect can transfer to the untrained tests.

Fine-tuning on the ASP

Training set:

Test

- Degree ordering
- Dimension mismatch
- Argument mismatch

Testing set:

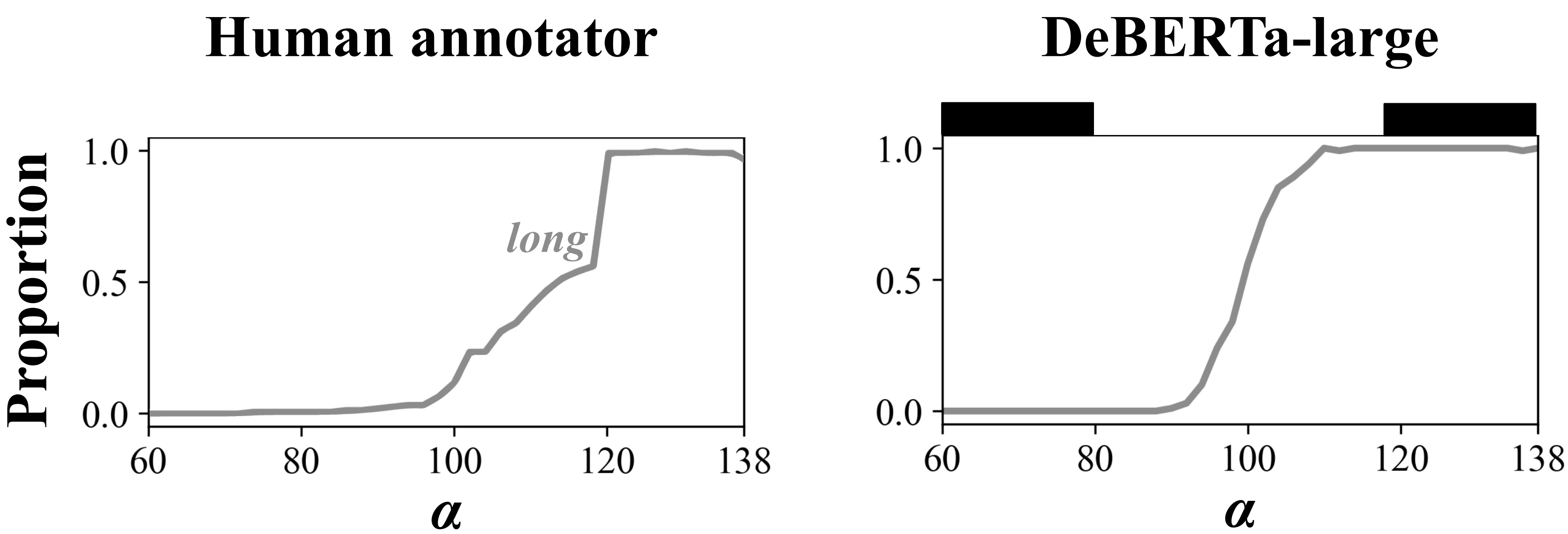
Test

- Booster
- Diminisher
- Negation
- Comparative
- Superlative

Setting of fine-tuning:

- Split the adjective vocabulary into training/testing set.
- Present the region <80 and >120 to models while fine-tuning.

Fine-tuning on the ASP

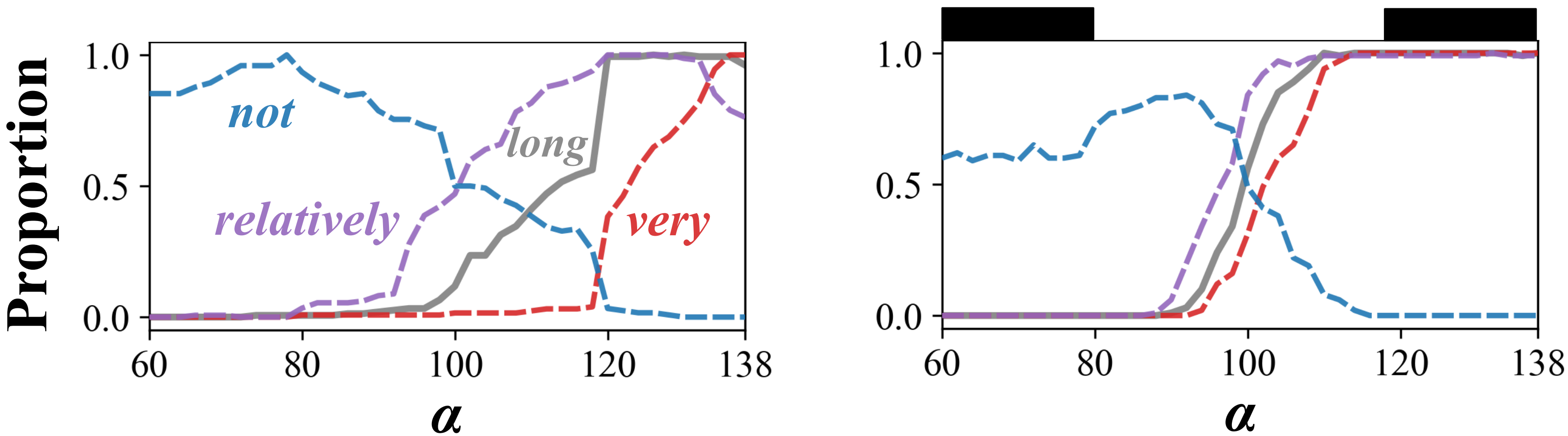


Premise: ***A 120 cm rod is long. A 80 cm rod is short.***

Hypotheses: ***A α cm rod is long*****. (training)** **$\alpha \in [60,138]$**

Fine-tuning on the ASP

Human annotator | **DeBERTa-large**



Premise: ***A 120 cm rod is long. A 80 cm rod is short.***

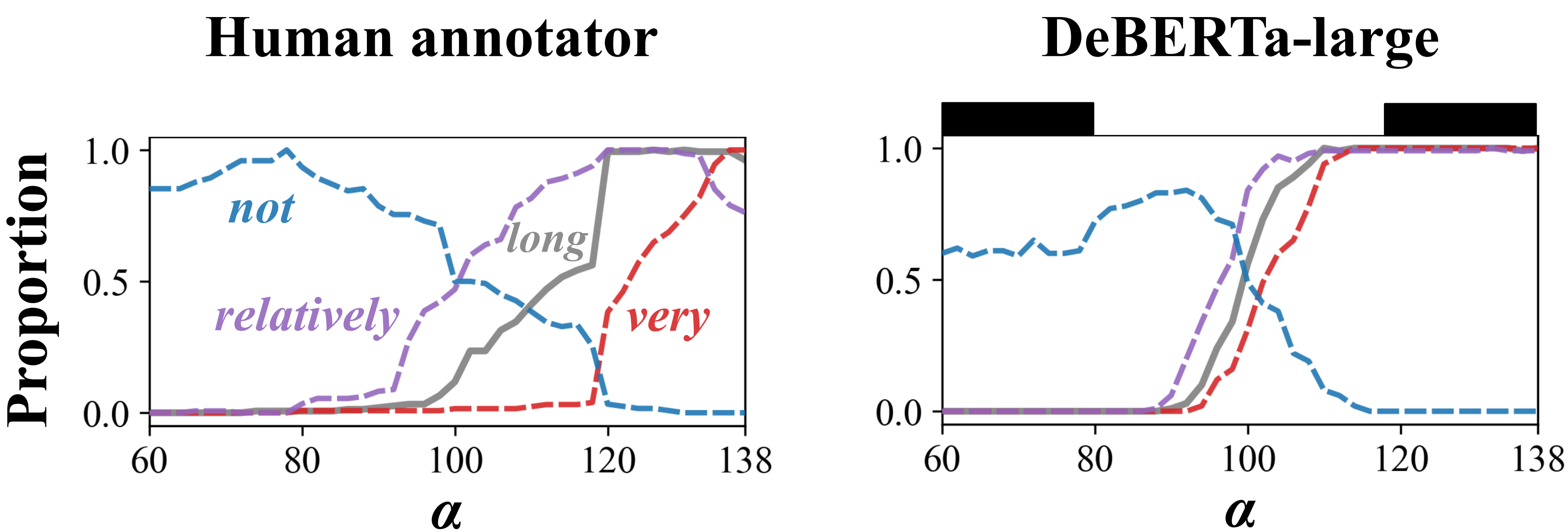
Hypotheses: ***A α cm rod is long.*** **(training)** $\alpha \in [60,138]$

A α cm rod is very long.

A α cm rod is relatively long.

A α cm rod is not long.

Fine-tuning on the ASP



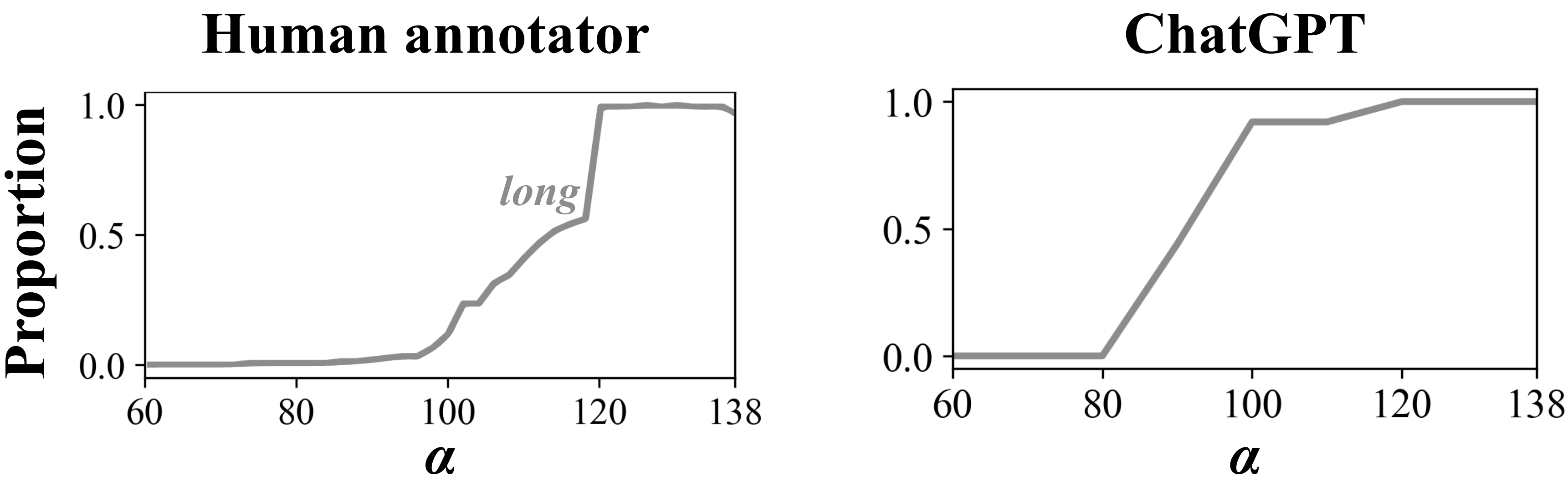
Language models generalized to the untrained tests.

⇓

✓ **b) Language models encode the degree semantics, but fail to apply it in the current task.**

LI A 80 cm rod is short. A 120 cm rod is long. Can we infer that "A 60 cm rod is long"? Please only answer "yes" or "no".

No.

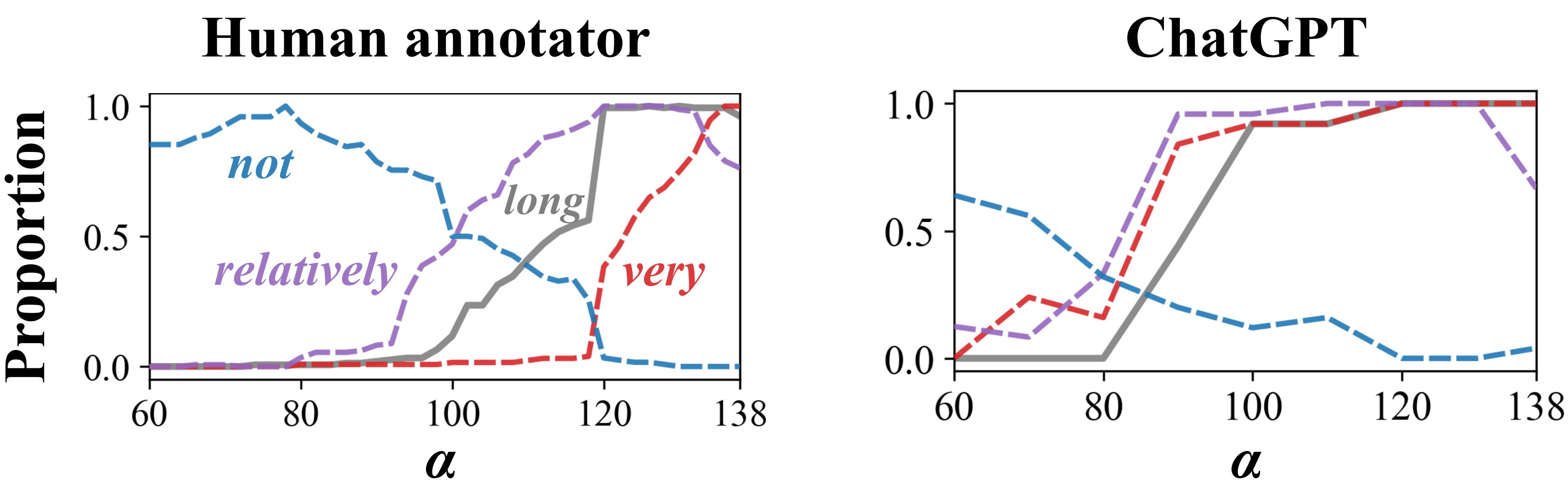


ChatGPT showed human-like behavior on the bare adjectives.

ChatGPT

A 80 cm rod is short. A 120 cm rod is long. Can we infer that "A $\boxed{60}$ cm rod is long"? Please only answer "yes" or "no".

No.



ChatGPT failed to modify the degree via the adverbs.

Conclusion

1. Language models fail to understand the degree semantics, which is the basic component of the semantics of adjectives.

2. With simply fine-tuning, language models can generalize the learning outcome to untrained tests, indicating the models can encode degree semantics.

Thanks!