Appendix A: Plausibility Control

When constructing the ASP dataset, to ensure the sentences we constructed conform to common world knowledge and natural usage, we semi-automatically generated the vocabulary items through the following process, with additional manual inspection.

- **Adjective collection.** Our ASP dataset involved adjectives from different semantic class. The relative, minimum, and maximum adjectives we used were all attested examples from Solt (2012). For extreme adjectives (used in the Degree ordering- entailment inference test), we collected them from Wilkinson and Tim (2016).
- **Dependency parsing.** We parsed Wikipedia based on dependency relation, and combined each word with its adjacent nodes as a bigram. We counted co-occurrence frequency for each bigram. For the adjective in classified adjective list, we ranked the nouns which formed bigram with the adjective based on co-occurrence frequency. We kept the top 300 nouns with highest co-occurrence frequency as candidates for each adjective.
- **Perplexity.** We computed the perplexity of the adjective-noun bigram to filter the uncommon combination. We translated each adjective-noun bigram to a simple sentence via the template *Noun is Adjective*, and used the pseudo-perplexity (Salazar et al. 2020) of the sentence to approximate the perplexity of the adjective-noun bigram. We computed the pseudo-perplexity via pre-trained RoBERTa-large following the implementation in Salazar et al. (2020). We took the top 50 nouns with the lowest perplexity as the final noun list for each adjective.
- **Distribution over Quantity (DoQ).** In the degree estimation task, we also needed to consider the plausibility of the combination between numerals and nouns on each given dimension. DoQ (Elazar et al. 2019) provided the multiple quantiles for certain physical dimension of nouns. We took the medians on the physical dimension of nouns as the typical numeral Num used for the construction of each test in the degree estimation task. The test range and interval of α was determined by the typical numeral Num: $\alpha \in$ [0.6×Num, 1.38×Num] for each noun, with a fixed interval of 0.02×Num. We rounded all numerals, and kept them as positive integers within the range of 0 to 20,000 by selecting the measurement units and nouns.

Appendix B: Human Annotation for the Degree Estimation Task

For the degree estimation task, the entailment ranges in Table 2 are determined based on human annotation. We randomly sample 10 adjective-noun pairs from each test implemented in the degree estimation task (except for the Dimension mismatch and the Argument mismatch test), and recast the NLI-style samples to a human cloze task. Examples of the human cloze task is listed in Appendix file (human/cloze.csv). Human annotators (N=40, 20~49 years old, mean age, 32.1 years; 29 females; all native English speakers) are presented with cloze questions, and are required to fill in the blank with a certain numeral range, based on the context provided to them. After normalizing the results of each annotator to a certain numeral range, we calculate the proportion of each normalized point selected by human, as the empirical scales of human annotators. The empirical scales are used to fit the threshold (i.e., δ in Table 2) of each test. We do not collect human annotations for the Dimension mismatch and the Argument mismatch tests, since the expected outcome for these tests are relatively clear.

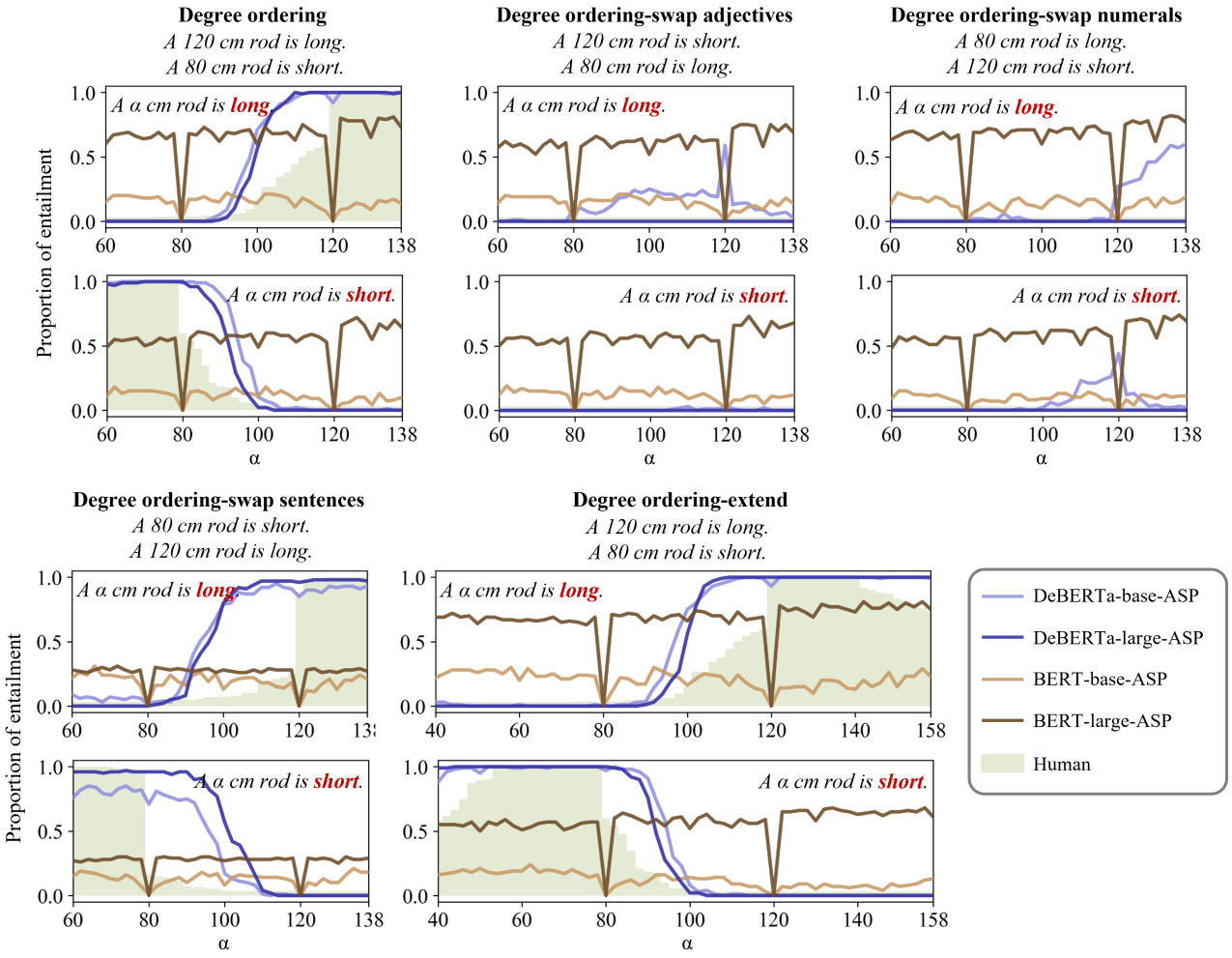
Appendix C: Fine-tuning and Inference Procedures on NLI

For NLI models, each sentence pair from the NLI datasets was concatenated in the following form: [CLS, premise, SEP, hypothesis, SEP], then the sequence was input to NLI models for encoding. We fed the CLS token into a 3-way softmax classifier to identify the entailment relations between premises and hypothesizes (entailment, neutral, or contradiction). The fine-tuning parameters for MNLI dataset were adopted from previous studies (Devlin et al. 2019; He, Gao, and Chen 2021) (shown in Appendix Table 1). We used the pre-trained models provided by huggingface (Wolf et al. 2020).

For zero-shot models, we translated NLI-style sentence pairs into human-readable form sentences by prompt, then

NLI dataset	MNLI	MNLI	MNLI	MNLI	/	/
Train model	BERT-base	BERT-large	DeBERTa-base	DeBERTa-large	T0 3B	T0 pp
Learning rate	2e-5	2e-5	2e-5	5e-6	/	/
Train epochs	3	2	2	2	/	/
Batch size	32	32	64	32	/	/
Weight decay	0.01	0.01	0.0	0.0	/	/
Warming steps	2000	2000	2000	160	/	/
Test accuracy (-m/-mm)	84.0/83.9	87.8/87.6	90.2/90.7	91.6/91.8	52.8/54.5	63.0/63.7
Accuracy in Stress Test-NR	36.6	64.0	67.1	72.3	29.3	40.1

Appendix Table 1. The fine-tuning parameters of MNLI dataset. The test accuracy of MNLI-dev and Stress Test Numerical Reasoning is shown in the last two rows.



Appendix Figure 1. Model performance on the control experiment.

input the readable sentences to zero-shot models (Sanh et al. 2021). The models generated the target choice based on the context, then mapped the target choice to NLI labels through a verbalizer. We used multiple 3-label NLI prompts for zero-shot models, and applied the majority vote among all prompts to get final predication. Appendix Table 1 listed the performance of zero-shot models, on the MNLI and Stress Test-Numerical reasoning dataset (Naik et al. 2018).

NLI dataset	ASP	ASP	ASP	ASP
Train model	BERT-base	BERT-large	DeBERTa-base	DeBERTa-large
Learning rate	2e-5	5e-6	2e-5	5e-6
Weight decay	0.0	0.0	0.0	0.0
Train epochs	{**12**,6,3}	{12,**6**,3}	{10,5,**3**}	{10,5,**3**}
Batch size	{1,**8**,32}	{1,**8**,32}	{1,8,**32**}	{1,8,**32**}
Warming steps	{**2000**,150}	{2000,**150**}	{2000,**150**}	{2000,**150**}

Appendix Table 2. The fine-tuning parameters on the ASP dataset. We perform grid-search for parameters. The final parameters used in our study are shown in bold.

Appendix D: Control Experiment for Models Fine-tuned on the ASP

We constructed a control experiment to exclude the possible heuristics might be learned by models during the fine-tuning on the ASP. We took *Degree ordering*-degree estimation test as a case study. For each premise, we swapped the numerals, adjectives, and sentences, respectively. By swapping numerals and adjectives, the theoretical scales in the premises were conflicted. Models should not infer the scale of adjectives due to the conflict, if these models truly understood the adjective scales. In the condition of swapping sentences, models should perform similar to the original *Degree ordering* test if these models did not make inference via simple heuristic of position embedding. Besides, we extended the original range of α in the condition of "-extend", while remaining other setting unchanged. The extended range could answer whether models understand the ordering relation between scales and numerals, rather than being over-fitted in

Type	Premise	Hypothesis	Entailment condition
Degree ordering*	The $\underline{\text{N}}$ is $\underline{\text{A}}$.	The $\underline{\text{N}}$ is extreme($\underline{\text{A}}$).	Extreme(A) → A
Dimension mismatch*	The $\underline{\text{N}}$ is $\underline{\text{A}}_1$.	The $\underline{\text{N}}$ is extreme($\underline{\text{A}}_2$).	∅
Argument mismatch*	The $\underline{\text{N}}_1$ is $\underline{\text{A}}$.	The $\underline{\text{N}}_2$ is $\underline{\text{A}}$.	∅
Booster*	The $\underline{\text{N}}$ is $\underline{\text{A}}$.	The $\underline{\text{N}}$ is very $\underline{\text{A}}$.	Very A → A
Diminisher*	The $\underline{\text{N}}$ is $\underline{\text{A}}$.	The $\underline{\text{N}}$ is relatively $\underline{\text{A}}$.	A → relatively A
Negation*	The $\underline{\text{N}}$ is $\underline{\text{A}}$.	The $\underline{\text{N}}$ is not antonym($\underline{\text{A}}$).	Relative adj. : antonym($\underline{\text{A}}$) → A Min. or max. adj. : antonym($\underline{\text{A}}$) ↔ A
Comparative	The $\underline{\text{N}}_1$ is comparative($\underline{\text{A}}$) than the $\underline{\text{N}}_2$.	(a) The $\underline{\text{N}}_1$ is $\underline{\text{A}}$. (b) The $\underline{\text{N}}_2$ is not $\underline{\text{A}}$.	Relative adj. : ∅ Minimum adj. : (a) Maximum adj. : (b)
Superlative*	The $\underline{\text{N}}$ is the super($\underline{\text{A}}$) $\underline{\text{N}}$ I have ever seen.	The $\underline{\text{N}}$ is the super($\underline{\text{A}}$) $\underline{\text{N}}$ in the world.	in the world → I have ever seen

Appendix Table 3. Templates for the entailment inference task. Task with "*" means the premise-hypothesis pair is exchangeable in this test. N stands for a noun. A stands for an adjective. Antonym() denotes that the antonym of an adjective. Extreme(), comparative(), and super() denotes the extreme, comparative, and superlative form of an adjective, respectively.

the numeral range of training set.

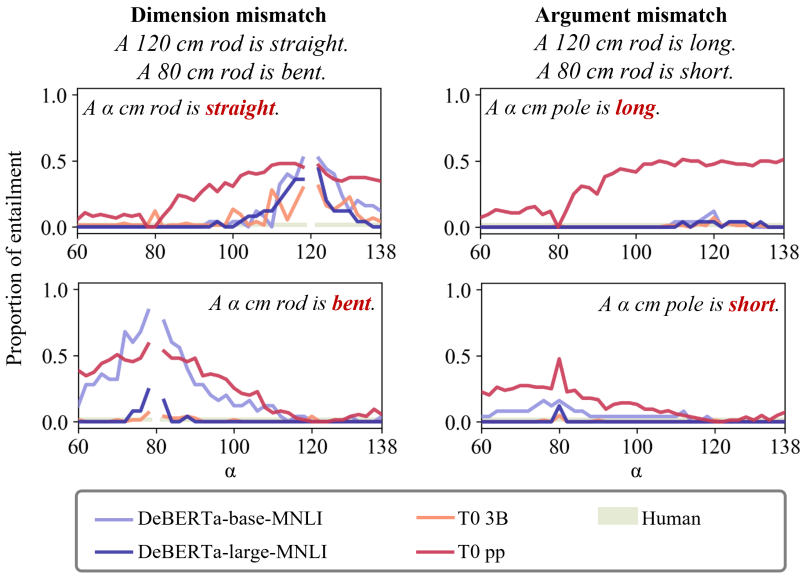
The evaluation result of control experiment showed that (see Appendix Figure 1), DeBERTa model refused to impose the ordering relation when the conflict appeared in the premises (swap adjectives or swap numerals), and performed similar to the *Degree ordering* test when the sentences in the premise were exchanged. Furthermore, DeBERTa models generalized the entailment pattern to the extended numeral range. The control experiment indicated that, after fine-tuning on the ASP, DeBERTa models made inference by correct semantic properties of adjectives, rather than superficial heuristic.

Appendix E: Numeral Range Manipulation for Models Fine-tuned on the ASP

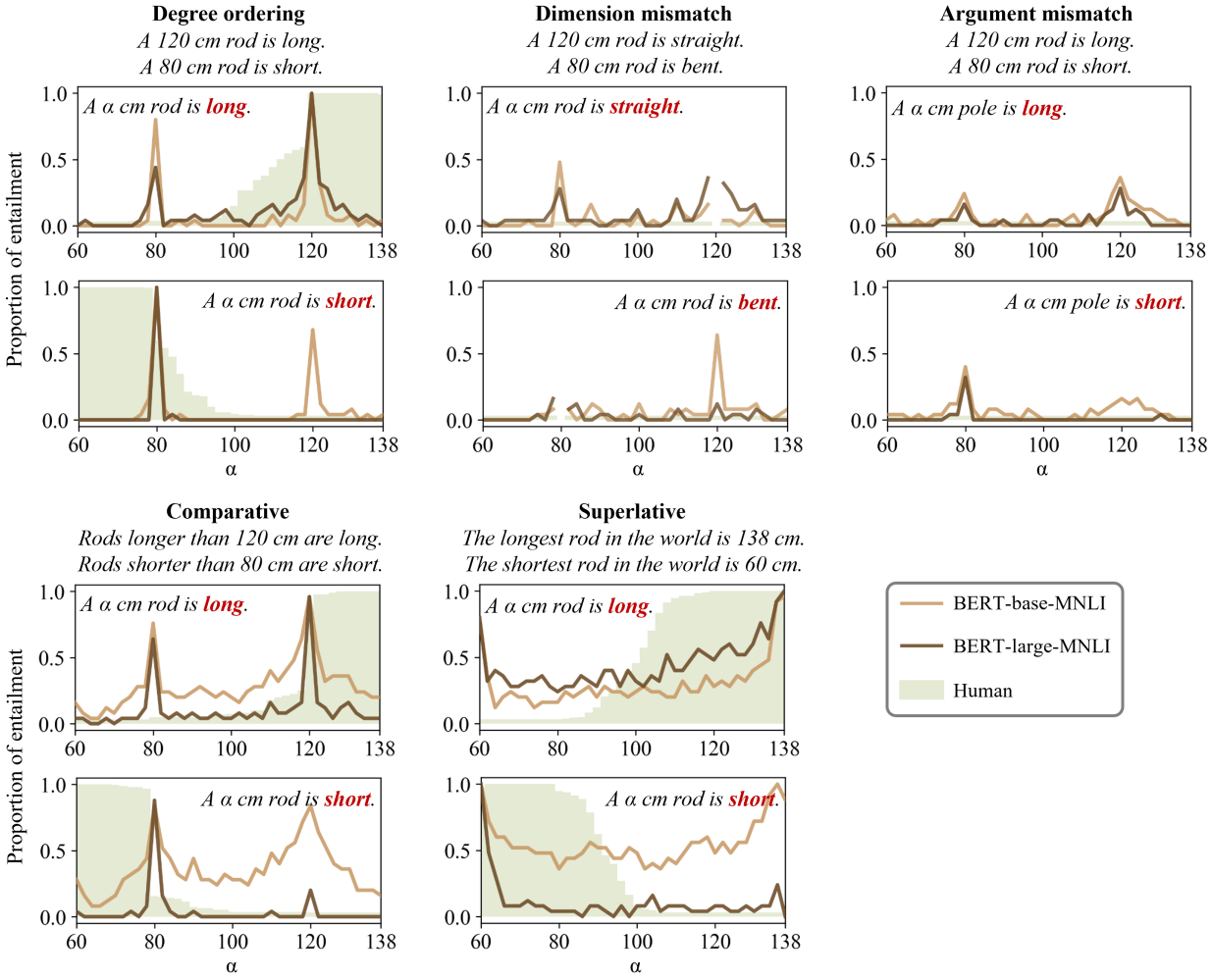
We manipulated the numerals in the premise of the degree estimation task to evaluate the robustness of models. The start and end of the range were randomly sampled. Then we equally split the range into 80 parts to match our previous experimental setup. For instance, after sampling, the range of α may be $[150, 545]$, with a fixed interval 5. The hypotheses are entailed by the premise only when $250 \leq \alpha \leq 450$. The evaluation result (Appendix Figure 8 and 9) were similar to the previous results in the main text.

Type	Premise	Hypothesis	Entailment condition
Degree ordering	A $\underline{\text{Num}_1}$ cm $\underline{\text{N}}$ is $\underline{\text{Pos}}$. A $\underline{\text{Num}_2}$ cm $\underline{\text{N}}$ is $\underline{\text{Neg}}$.	A $\underline{\alpha}$ cm $\underline{\text{N}}$ is $\underline{\text{Pos}}$ A $\underline{\alpha}$ cm $\underline{\text{N}}$ is $\underline{\text{Neg}}$.	$\alpha \geq \text{Num}_1 - \delta$ $\alpha \leq \text{Num}_2 + \delta$
Dimension mismatch	A $\underline{\text{Num}_1}$ cm $\underline{\text{N}}$ is $\underline{\text{Pos}_{\text{ur}}}$. A $\underline{\text{Num}_2}$ cm $\underline{\text{N}}$ is $\underline{\text{Neg}_{\text{ur}}}$.	A $\underline{\alpha}$ cm $\underline{\text{N}}$ is $\underline{\text{Pos}_{\text{ur}}}$ A $\underline{\alpha}$ cm $\underline{\text{N}}$ is $\underline{\text{Neg}_{\text{ur}}}$.	$\emptyset$ ($\alpha \neq \text{Num}_1$) $\emptyset$ ($\alpha \neq \text{Num}_2$)
Argument mismatch	A $\underline{\text{Num}_1}$ cm $\underline{\text{N}_1}$ is $\underline{\text{Pos}}$. A $\underline{\text{Num}_2}$ cm $\underline{\text{N}_1}$ is $\underline{\text{Neg}}$.	A $\underline{\alpha}$ cm $\underline{\text{N}_2}$ is $\underline{\text{Pos}}$ A $\underline{\alpha}$ cm $\underline{\text{N}_2}$ is $\underline{\text{Neg}}$.	$\emptyset$ $\emptyset$
Booster	A $\underline{\text{Num}_1}$ cm $\underline{\text{N}}$ is $\underline{\text{Pos}}$. A $\underline{\text{Num}_2}$ cm $\underline{\text{N}}$ is $\underline{\text{Neg}}$.	A $\underline{\alpha}$ cm $\underline{\text{N}}$ is very $\underline{\text{Pos}}$ A $\underline{\alpha}$ cm $\underline{\text{N}}$ is very $\underline{\text{Neg}}$.	$\alpha \geq \text{Num}_1 + \delta$ $\alpha \leq \text{Num}_2 - \delta$
Diminisher	A $\underline{\text{Num}_1}$ cm $\underline{\text{N}}$ is $\underline{\text{Pos}}$. A $\underline{\text{Num}_2}$ cm $\underline{\text{N}}$ is $\underline{\text{Neg}}$.	A $\underline{\alpha}$ cm $\underline{\text{N}}$ is relatively $\underline{\text{Pos}}$ A $\underline{\alpha}$ cm $\underline{\text{N}}$ is relatively $\underline{\text{Neg}}$.	$\alpha \geq \text{Num}_1 - \delta$ $\alpha \leq \text{Num}_2 + \delta$
Negation	A $\underline{\text{Num}_1}$ cm $\underline{\text{N}}$ is $\underline{\text{Pos}}$. A $\underline{\text{Num}_2}$ cm $\underline{\text{N}}$ is $\underline{\text{Neg}}$.	A $\underline{\alpha}$ cm $\underline{\text{N}}$ is not $\underline{\text{Pos}}$ A $\underline{\alpha}$ cm $\underline{\text{N}}$ is not $\underline{\text{Neg}}$.	$\alpha \leq \text{Num}_2 + \delta$ $\alpha \geq \text{Num}_1 - \delta$
Comparative	$\underline{\text{N}}$ comparative($\underline{\text{Pos}}$) than $\underline{\text{Num}_1}$ cm is $\underline{\text{Pos}}$. $\underline{\text{N}}$ comparative($\underline{\text{Neg}}$) than $\underline{\text{Num}_2}$ cm is $\underline{\text{Neg}}$.	A $\underline{\alpha}$ cm $\underline{\text{N}}$ is $\underline{\text{Pos}}$ A $\underline{\alpha}$ cm $\underline{\text{N}}$ is $\underline{\text{Neg}}$.	$\alpha \geq \text{Num}_1$ $\alpha \leq \text{Num}_2$
Superlative	The super($\underline{\text{Pos}}$) $\underline{\text{N}}$ in the world is $\underline{\text{UB}}$ cm. The super($\underline{\text{Neg}}$) $\underline{\text{N}}$ in the world is $\underline{\text{LB}}$ cm.	A $\underline{\alpha}$ cm $\underline{\text{N}}$ is $\underline{\text{Pos}}$ A $\underline{\alpha}$ cm $\underline{\text{N}}$ is $\underline{\text{Neg}}$.	$\alpha \geq \text{UB} - \delta$ $\alpha \leq \text{LB} + \delta$

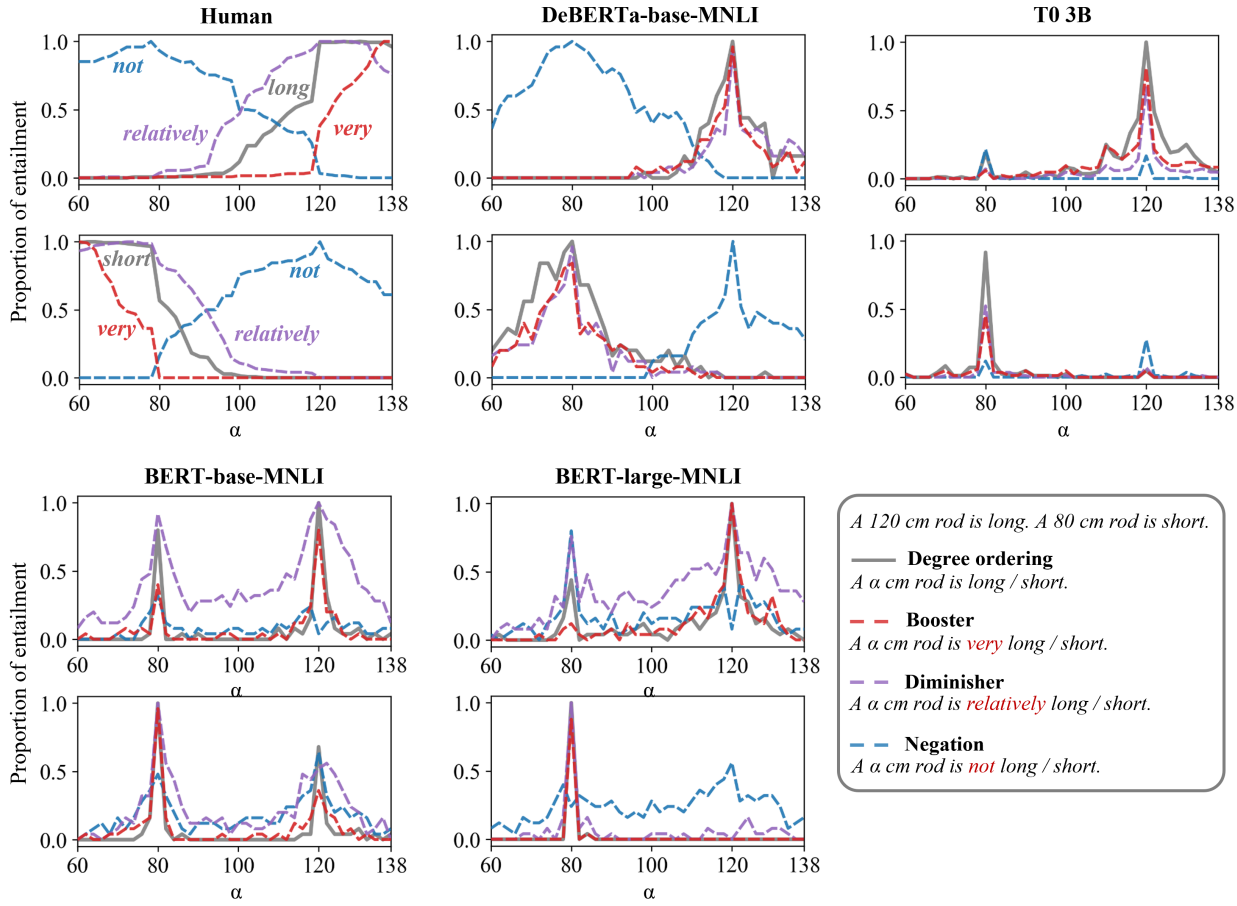
Appendix Table 4. Templates for the degree estimation task. Num_1 and Num_2 are numerals derived from typical numerals Num, with different magnification (×1.2 and ×0.8, respectively). UB and LB stands for the upper and lower boundary of the range of α (i.e., 1.38×Num and 0.6×Num, respectively). Pos and Neg stand for the positive and negative adjective from the same antonym pair. The subscript ur means the adjective is irrelevant to the physical dimension.



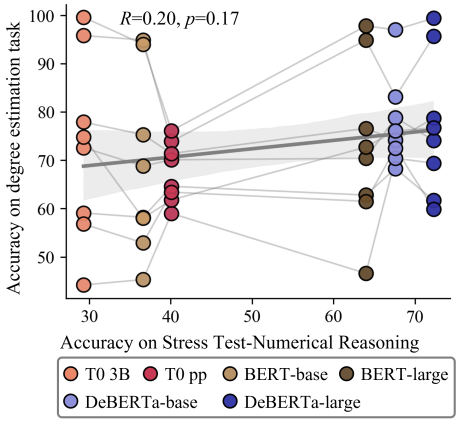
Appendix Figure 2. Model performance on the *Dimension mismatch* and *Argument mismatch* tests of the degree estimation task. For the *Dimension mismatch* test, we discard the hypothesis which is completely overlapped with the premise (e.g., α=120 for positive adjectives).



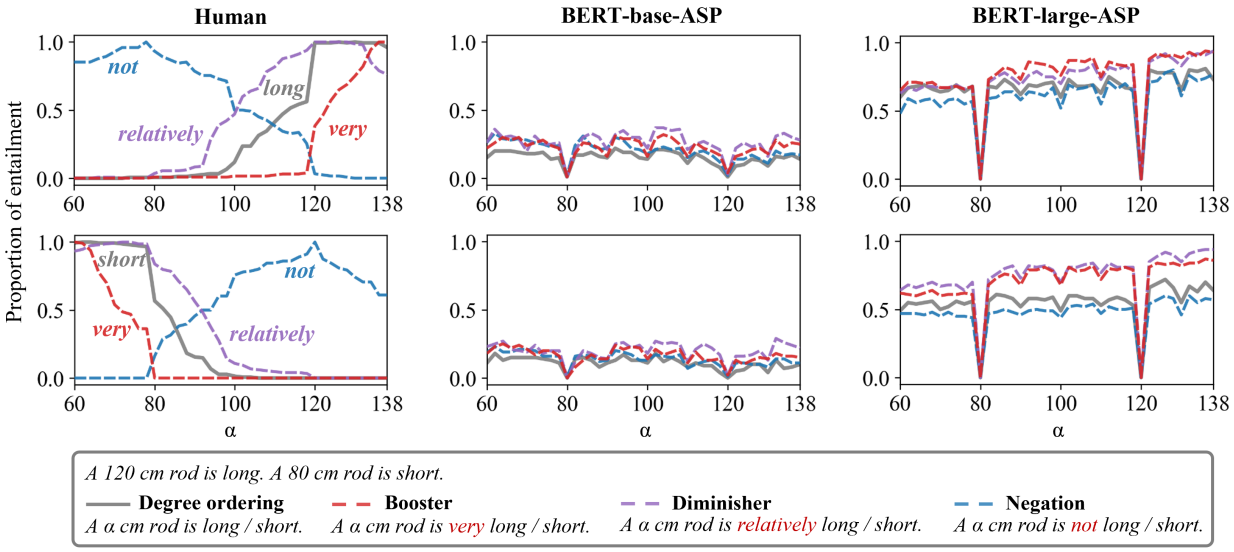
Appendix Figure 3. Model performance on the the degree estimation task.



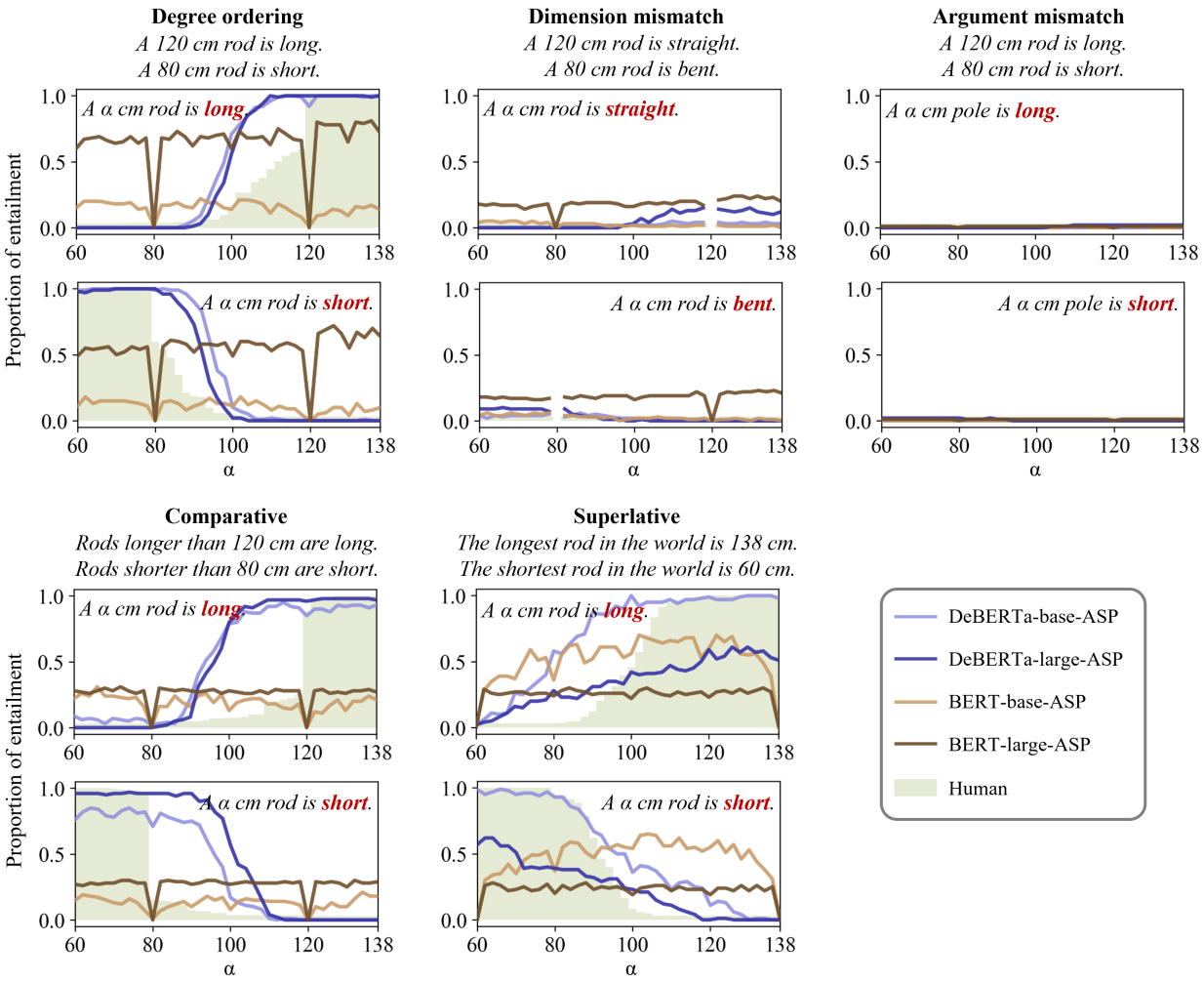
Appendix Figure 4. Model performance on the *Booster*, *Diminisher*, and *Negation* tests of the degree estimation task.



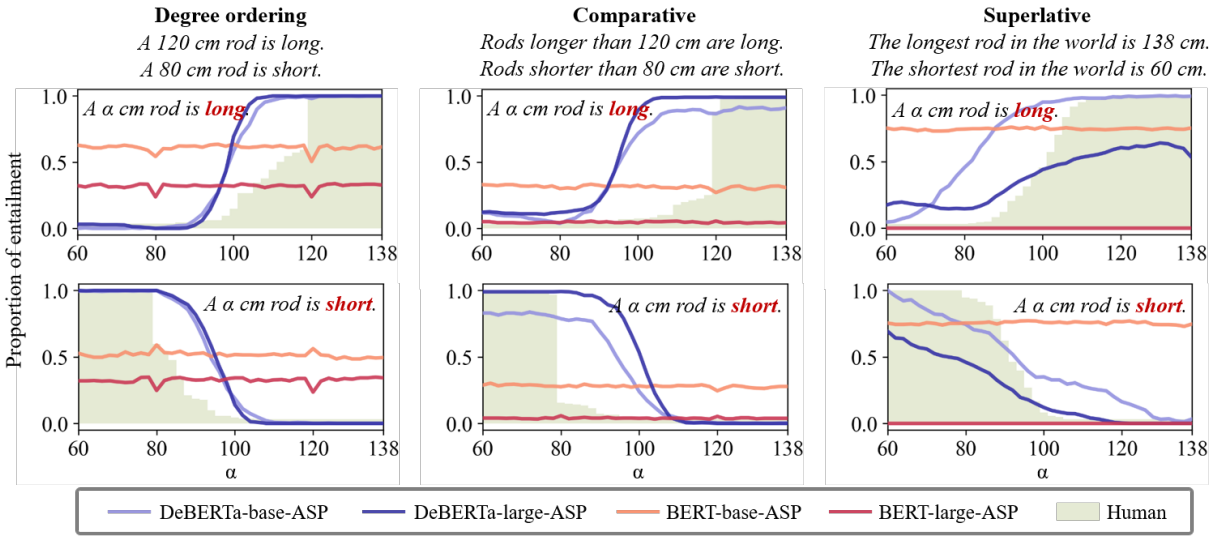
Appendix Figure 5. Correlation between the performance on the degree estimation task and the numerical reasoning ability. We use the accuracy on the Stress test-Numerical reasoning set as the metric for numerical reasoning ability. Each dot stands for the model performance on certain test of the degree estimation task. The dots for the same degree estimation tests are connected by light-grey lines. The line shadow represents the confidence interval (95%) of the regression line.



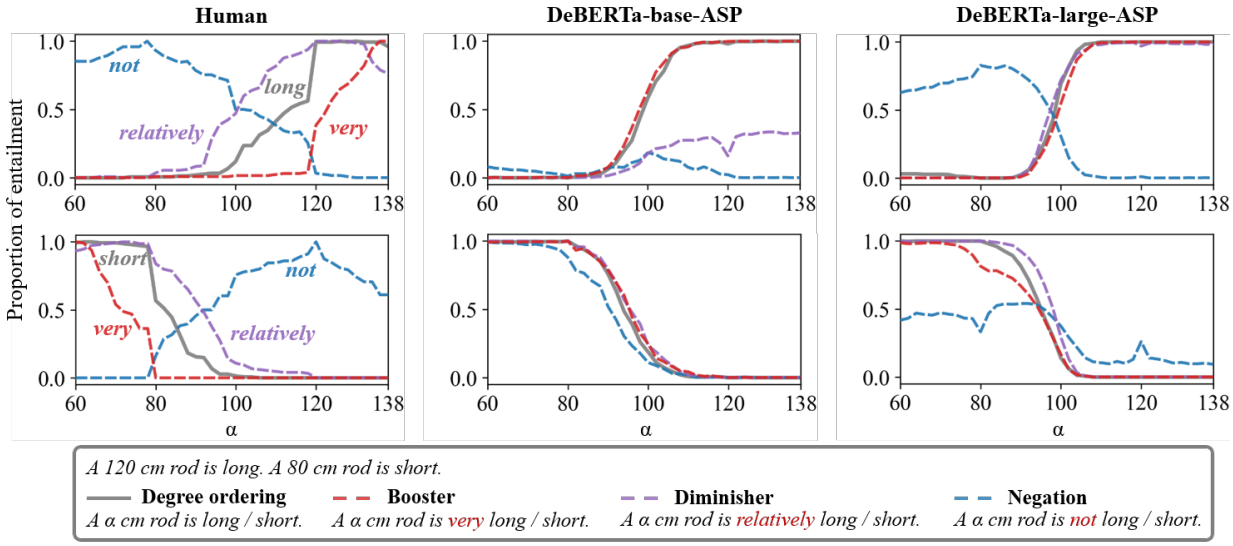
Appendix Figure 6. Model performance on the *Booster*, *Diminisher*, and *Negation* tests of the degree estimation task.



Appendix Figure 7. Model performance on the degree estimation task.



Appendix Figure 8. Model performance on the *Booster*, *Diminisher*, and *Negation* tests of the degree estimation task, after manipulating the range.



Appendix Figure 9. Model performance on the degree estimation task, after manipulating the range.

References

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 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)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Elazar, Y.; Mahabal, A.; Ramachandran, D.; Bedrax-Weiss, T.; and Roth, D. 2019. How Large Are Lions? Inducing Distributions over Quantitative Attributes. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3973–3983. Florence, Italy: Association for Computational Linguistics.

He, P.; Gao, J.; and Chen, W. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Naik, A.; Ravichander, A.; Sadeh, N.; Rose, C.; and Neubig, G. 2018. Stress Test Evaluation for Natural Language Inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, 2340–2353. Santa Fe, New Mexico, USA: Association for Computational Linguistics.

Salazar, J.; Liang, D.; Nguyen, T. Q.; and Kirchhoff, K. 2020. Masked Language Model Scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2699–2712. Online: Association for Computational Linguistics.

Sanh, V.; Webson, A.; Raffel, C.; Bach, S. H.; Sutawika, L.; Alyafeai, Z.; Chaffin, A.; Stiegler, A.; Scao, T. L.; Raja, A.; et al. 2021. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*.

Solt, S. 2012. Comparison to arbitrary standards. In *Proceedings of Sinn und Bedeutung*, volume 16, 557–570.

Wilkinson, B.; and Tim, O. 2016. A Gold Standard for Scalar Adjectives. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 2669–2675. Portorož, Slovenia: European Language Resources Association (ELRA).

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; Davison, J.; Shleifer, S.; von Platen, P.; Ma, C.; Jernite, Y.; Plu, J.; Xu, C.; Scao, T. L.; Gugger, S.; Drame, M.; Lhoest, Q.; and Rush, A. M. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45. Online: Association for Computational Linguistics.