MASP: A Multilingual Dataset for Probing Scalar Modifier Understanding in LLMs

Xinyu Gao[1], Nai Ding[1], Wei Liu[1*]
[1]College of Biomedical Engineering and Instrument Sciences, Zhejiang University, Hangzhou, China
{xinyu_gao, ding_nai, liuweizju}@zju.edu.cn

Abstract

This study aims to test how large language models (LLMs) understand gradable adjectives and whether their understanding compares with humans, under the framework of formal semantics. We introduce a diagnostic dataset, referred to as the Modifier-Adjective Scale Probe (MASP), to evaluate how well LLMs understand a gradable adjective (e.g., *long*) when the adjective is combined with one modifier (e.g., *very long* or *slightly long*, a condition referred to as *degree modification*) or is further negated (e.g., *very not long* and *not very long*, a condition referred to as *compositional negation*). The dataset consists of over 80,000 natural language inference questions in both Chinese and English. We apply the MASP dataset to test both humans and 11 popular LLMs, including GPT-4o and Gemini-2.0-Flash. The results show that most LLMs can correctly understand whether a modifier boosts (e.g., *very*) an adjective. However, they fail to understand the modifiers that weaken the degree and the negation forms of modifiers. Furthermore, we parameterize the human and LLM behavior, and find that the judgment patterns of LLMs differ from humans especially in the Chinese tests. These findings suggest that LLMs are still not well aligned with humans in terms of the interpretation of simple adjective phrases, and MASP provides a new approach to quantify the interpretation of adjective phrases in LLMs.

1 Introduction

The capacity of large language models (LLMs) to encode semantic information has been extensively studied in many linguistic tasks such as natural language inference (NLI) (Bowman et al., 2015; Devlin et al., 2019; Ruis et al., 2023). However, the semantics of certain linguistic expressions can vary in degree, depending on modifiers. For instance, the proposition "*John is tall*" may hold true when assessed relative to middle school students. However, the proposition "*John is extremely tall*" may prove false because the modifier "*extremely*" substantially boosts the degree requirement for tallness, exceeding John's actual height qualification. While formal semantics provides the theoretical frameworks for describing the semantics of degree (Kennedy, 2007; Lassiter, 2017), it remains debated whether LLMs can encode degree semantic information (Liu et al., 2023a; Lin et al., 2024), especially when gradable adjectives/verbs are combined with scalar modifiers. LLMs do not explicitly encode the degree semantic information, but the absence of explicit encoding does not necessarily imply that the information is entirely unnecessary for LLMs. One possibility is that degree semantics can functionally explain the linguistic behavior of LLMs. Here, we investigate whether LLMs can encode degree semantic information, by quantitatively examining LLMs' understanding of gradable adjectives with one modifier (e.g., *very long* and *slightly long*) or are further negated (e.g., *very not long* and *not very long*).

We build a diagnostic dataset, referred to as the Modifier-Adjective Scale Probe (MASP), to evaluate LLMs' understanding of the gradable adjectives and scalar modifiers. MASP adopts the Natural Language Inference framework, tasking LLMs with determining entailment relationships in sentence pairs (Nie et al., 2020). In recent years, various NLI-style diagnostic datasets have been developed for

[*]Corresponding author: Wei Liu

Proceedings of the 24th China National Conference on Computational Linguistics, pages 1003-1019, Jinan, China, August 11-14, 2025.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

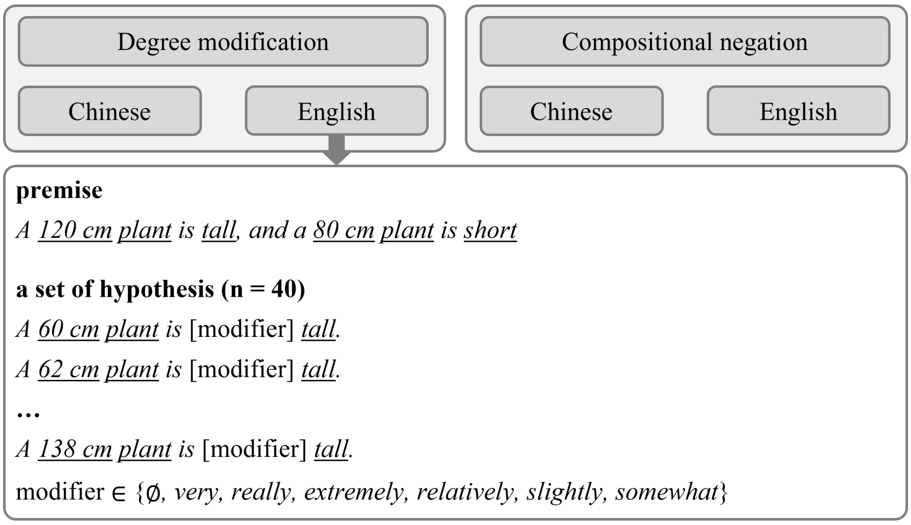

Degree modification
Chinese
English
Compositional negation
Chinese
English
premise
A 120 cm plant is tall, and a 80 cm plant is short
a set of hypothesis (n = 40)
A 60 cm plant is [modifier] tall.
A 62 cm plant is [modifier] tall.
…
A 138 cm plant is [modifier] tall.
modifier ∈ {∅, very, really, extremely, relatively, slightly, somewhat}

Figure 1: The overview of the MASP dataset. The dataset is organized along two conditions, with each condition instantiated in both Chinese and English tests. Underlined words in the figure indicate those inserted into templates.

assessing targeted linguistic competencies, notably lexical semantic inference (Sinha et al., 2019) and commonsense inference (Tafjord et al., 2022), and to probe the superficial heuristic learned by LLMs (McCoy et al., 2019). Some studies also focused on the degree semantics of bare adjectives (Liu et al., 2023b), but they did not analyze the compositionality between adjectives and modifiers, nor did they explore the cross-linguistic commonalities of degree semantics in LLMs. To address this, MASP contains both Chinese and English tests, and separately evaluate 2 main aspects of adjective and modifier interpretation, i.e., *degree modification* and *compositional negation* (Figure 1).

We apply the MASP dataset to evaluate 11 state-of-the-art LLMs (e.g., Gemini-2.0-Flash and GPT-4o, all GPT-4o results reported in this paper are based on the gpt-4o-2024-08-06 model release.), and found that most LLMs can capture the variations when a modifier boosts (e.g., *very*) an adjective, but fail to understand the modifiers that weaken the degree and the negation forms of modifiers. LLMs with better language comprehension capabilities encode degree semantics that are more similar to those of humans. In addition, we parameterize the human and LLM behavior using three factors, including polarity confidence, decision thresholds and judgment precision, and find that humans are more consistent and rapid in adjusting their judgments as the modification strength of modifiers changes. The main contributions of our study are: (i) constructing a theoretically motivated NLI diagnostic dataset to examine the LLMs' understanding of compositional scalar expressions, (ii) demonstrating that LLMs encode degree semantic information differently from humans, exhibiting lower consistency in responses to modifiers and slower adaptation to increased modification strength.

2 Dataset Construction

2.1 Task

The dataset is organized along two conditions: (1) *degree modification*, and (2) *compositional negation*, with each condition instantiated in both Chinese and English tests. For each test, we design a degree estimation task following the adjective probing framework described by Liu et al. (2023b). The degree estimation task requires LLMs to assess whether the premise entails or fails to entail each hypothesis from the set of hypotheses separately. Each premise contains antonym pairs and is evaluated against 40 templated hypotheses exhibiting exclusive numerical variation. These numeral values are systematically sampled at fixed intervals from a predefined range (see the examples in Figure 1), with all premises and hypotheses are generated through template-based construction.

Degree modification	premise	hypothesis	entailment range
Chinese	120 cm 的植物是高的, 80 cm 的植物的矮的。	α cm 的植物是[modifier] 高的 α cm 的植物是[modifier] 矮的	$\alpha > 120 - \delta$ $\alpha > 80 + \delta$
English	A 120 cm plant is long, and a 80 cm plant is short.	A α cm plant is [modifier] tall. A α cm plant is [modifier] short.	$\alpha > 120 - \delta$ $\alpha < 80 + \delta$

Compositional negation	premise	hypothesis	entailment range
Chinese	120 cm 的植物是高的, 80 cm 的植物的矮的。	α cm 的植物是不 [modifier] 高的 α cm 的植物是[modifier] 不高的 α cm 的植物是不 [modifier] 矮的 α cm 的植物是[modifier] 不矮的	$\alpha < 80 + \delta$ $\alpha < 80 + \delta$ $\alpha > 120 - \delta$ $\alpha > 120 - \delta$
English	A 120 cm plant is long, and a 80 cm plant is short.	A α cm plant is **not** [modifier] tall. A α cm plant is [modifier] **not** tall. A α cm plant is **not** [modifier] short. A α cm plant is [modifier] **not** short.	$\alpha < 80 + \delta$ $\alpha < 80 + \delta$ $\alpha > 120 - \delta$ $\alpha > 120 - \delta$

Table 1: Examples for each condition. The placeholder [modifier] is the candidate scalar modifier that can be choice from $\{\emptyset$, *very*, *really*, *extremely*, *relatively*, *slightly*, *somewhat*$\}$. $\emptyset$ represents bare adjectives. The *entailment range* column specifies the entailment bounds of α relative to each sentence pair. In these examples, $\alpha \in$ [60,140], with a fixed interval of 2. δ is a value depending on the modifier (Solt, 2019; Kennedy and McNally, 2005).

The Chinese version of the tests is directly translated from the English version.To ensure semantic and pragmatic equivalence across languages, we carefully adjust the translations based on linguistic structure (e.g., word order). These adjustments include modifying word order, selecting appropriate scalar modifiers (e.g., rendering *slightly* as 稍微 or 有点), and handling differences in negation scope. All translated items are reviewed and refined by native Chinese speakers to ensure naturalness and consistency with everyday usage.

In total, we construct a total of 40,000 premise–hypothesis pairs for Chinese and English, respectively, covering 7 scalar modifiers under each condition (2,000 for each modifier in *degree modification* and 4,000 for each negated modifier in *compositional negation*, examples for each condition are shown in Table 1. In the following, we detail how the premises and hypotheses are constructed for each condition.

2.2 Degree Modification

Scalar modifiers modulate the internal thresholds of gradable adjectives by altering the degree of adjectives. Drawing from Kennedy's (2007) framework, we define a continuum of 7 scalar modifiers ordered by linguistic intuition:

$somewhat \rightarrow slightly \rightarrow relatively \rightarrow \emptyset \rightarrow really \rightarrow very \rightarrow extremely$

Here, the modification strength gradually increased from left to right, with the bare adjective ($\emptyset$) serving as the baseline. Modifiers to the left of the bare adjective are diminishers (e.g., *somewhat*, *slightly*), which weaken the degree of the adjective they combine with (Quirk et al., 1985). Modifiers to the right are intensifiers (e.g., *very*, *extremely*), which strengthen the scalar interpretation (Paradis, 2008). For both Chinese and English tests, each premise-hypothesis pair retains the identical adjective, and the hypothesis diverges from its premise through lexical modification (Table 1).

2.3 Compositional Negation

The degree of adjectives can be operated by the combination of modifiers. For instance, "*not very tall*" indicates the complement of "*very tall*", while "*very not tall*" indicates the enhanced degree of "*not tall*"

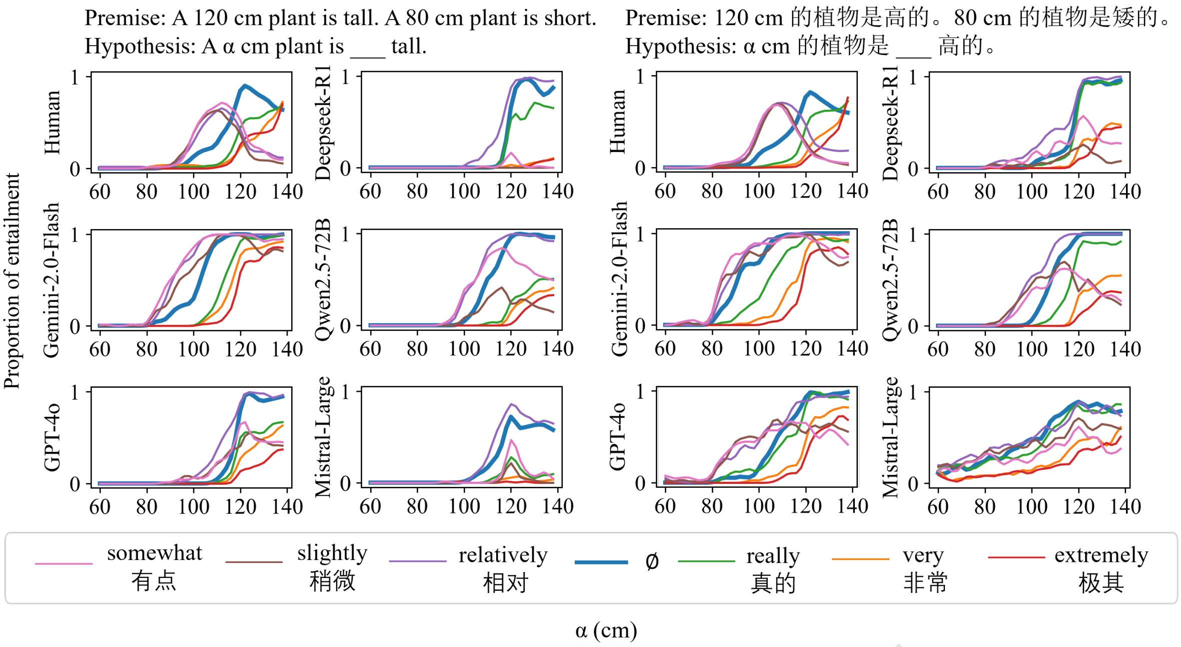


Figure 2: Human and LLM behavior on both Chinese and English tests of *degree modification* with positive adjectives (e.g., *tall*). See the complete results in Appendix Figures 1 and 2.

(Hersh and Caramazza, 1976). To investigate how the modifier interacts with others, we introduce two negation forms for each modifier (Table 1):

Modifier negation*: A α cm plant is not* [modifier] *tall.*

Adjective negation*: A α cm plant is* [modifier] *not tall.*

These forms test whether LLMs can correctly distinguish between modifier negation (reversing the effect of modifiers) and adjective negation (reversing the degree of adjectives). For both Chinese and English tests, each premise-hypothesis pair retains the identical adjective, and the hypothesis diverges from its premise through negation operations (modifier negation or adjective negation form) (Table 1).

3 Systematic Sensitivity to Scalar Modifiers

3.1 Experimental Setup

We evaluated 11 LLMs on MASP, including:

Open-Source Models: LLaMA-3.1-8B, LLaMA-3.1-70B, LLaMA-3.3-70B (Grattafiori et al., 2024); Mistral-7B (Jiang et al., 2023), Mistral-Small, Mistral-Large; Qwen2.5-7B, Qwen2.5-72B (Yang et al., 2024); DeepSeek-R1 (DeepSeek-AI et al., 2025);

Closed-Source Models: Gemini-2.0-Flash (Google DeepMind, 2025), GPT-4o (OpenAI et al., 2023).

LLMs were tested using a zero-shot NLI paradigm following the prompts described in a recent study (Reynolds and McDonell, 2021). Each model was instructed to identify the entailment relation between a premise and hypothesis without in-context examples (see the prompts in Appendix A).

To further ensure robustness, we conducted a consistency analysis using four prompt variants that differed in surface phrasing but conveyed the same NLI task. These included: (1) "Determine whether the premise entails the hypothesis", (2) "Does the premise entail the hypothesis? ", (3) "Is the premise true based on the hypothesis? " and (4) "Can the hypothesis be logically inferred from the premise?". We randomly sampled 500 pairs of sentences from the MASP dataset and evaluated five representative LLMs (GPT-4o, Gemini-2.0-Flash, DeepSeek-R1, LLaMA-3.3-70B and Mistral-Large). The agreement across prompt variants ranged from 0.970 to 0.992, with deviations within ±0.050 of the main results, indicating stable model behavior across different prompt wordings.

The closed-source models and DeepSeek-R1 were queried using the OpenRouter API, and other open-

Model	Chinese							English						
	∅	非常	真的	极其	相对	稍微	有点	∅	very	really	extremely	relatively	slightly	somewhat
Deepseek-R1	0.88	0.82	0.84	0.82	0.74	0.6	0.62	0.8	0.72	0.84	0.62	0.72	/	0.8
GPT-4o	**0.94**	0.9	**0.94**	0.90	0.74	0.74	0.78	0.87	**0.82**	0.86	0.84	0.74	0.7	0.87
Gemini-2.0-Flash	**0.94**	**0.93**	0.9	**0.92**	**0.82**	**0.80**	**0.82**	**0.96**	0.78	**0.92**	0.86	**0.84**	0.80	**0.96**
Llama-3.1-8b	0.68	0.32	0.43	0.36	0.33	0.45	0.38	0.92	0.8	0.86	0.82	0.75	0.73	0.92
Llama-3.1-70B	0.84	0.71	0.62	0.67	0.56	0.38	0.31	0.9	0.81	0.86	**0.88**	0.77	0.48	0.9
Llama-3.3-70B	0.9	0.7	0.72	0.72	0.68	0.62	0.56	0.9	0.79	0.85	0.8	0.61	/	0.9
Mistral-7B	0.81	0.82	0.8	0.65	0.42	0.56	0.42	0.94	0.8	0.88	0.76	0.77	0.68	0.94
Mistral-Small	0.82	0.53	0.64	0.64	0.54	0.58	0.5	0.84	0.74	0.9	0.74	0.8	**0.82**	0.84
Mistral-Large	0.76	0.53	0.64	0.64	0.53	0.58	0.5	0.88	0.54	0.74	/	0.68	0.23	0.88
Qwen2.5-7B	0.52	/	0.36	/	0.18	/	/	0.88	0.22	0.48	0.2	0.48	0.18	0.88
Qwen2.5-72B	0.75	0.8	0.83	0.75	0.78	0.62	0.57	0.88	/	/	/	0.71	/	0.88

Table 2: The Pearson correlations between human and LLM behavior on both Chinese and English tests of *degree modification*. Since the proportion of entailment obtained by some LLMs is 0, the correlation cannot be calculated and is marked as "/".

source models were run locally. Following previous studies (Brown et al., 2020), we configured temperature=0 to eliminate response stochasticity and max_tokens=200 to constrain generation length across all tests.

We also collected human annotation on the *degree modification* of MASP, involving 80 participants for Chinese and 40 for English. Each participant was required to annotate 80 questions. The human annotation for the degree estimation task was designed in a structured numerical reasoning format to assess degree semantic interpretations under scalar modifiers (see the instruction in Appendix B). For both human and LLM responses, we calculated the proportion of entailment – the number of responses labeled as "*entailment*" (or "蕴含" in Chinese) divided by the total number of responses. All human responses were pooled to calculate the proportion of entailment.

3.2 Result

3.2.1 Modification Strength of Single Modifier

To compare the modification strength between different modifiers, we analyzed the human and LLM behavior when adjectives were combined with different modifiers. This analysis focused on how the proportion of entailment of humans and LLMs varied with modifier-adjective combinations.

We visualized the behavior of humans and five representative LLMs in Figure 2, with the complete results presented in Appendix Figures 1 and 2. In both Chinese and English, the proportion of entailment shifted upwards on the scale for an intensifier (e.g., *extremely*) compared to the bare adjectives (Figure 2), indicating that both humans and LLMs applied stricter semantic thresholds when the strength of modification increased. Similar patterns emerged for negative adjectives (e.g., *short*) (Appendix Figures 3, 4, 5 and 6). These consistent patterns across Chinese and English suggested that humans and LLMs exhibited cross-linguistic understanding when adjectives were combined with intensifiers.

A divergence was observed between human and LLM behavior when adjectives were combined with diminishers (e.g., *relatively*). For humans, the proportion of entailment shifted downward compared to the bare adjectives, and peaked at mid-scale degrees (e.g., α = 110 cm), reflecting a bounded window for diminishers. In contrast, while the proportion of entailment also shifted downward along the scale, LLMs exhibited a monotonic increase rather than peaking at specific degrees, suggesting a failure to capture the attenuating semantics of diminishers.

These findings confirmed that humans and LLMs exhibited the cross-linguistic scalar sensitivity to degree modification, capturing the general direction and relative magnitude of graded modifiers. Nonetheless, divergences from human behavior also highlighted residual challenges in LLMs' understanding of degree semantics.

To ensure that these results are not artifacts of a particular model or run, we also assessed the consistency across models themselves. We computed Pearson correlations between five representative LLMs

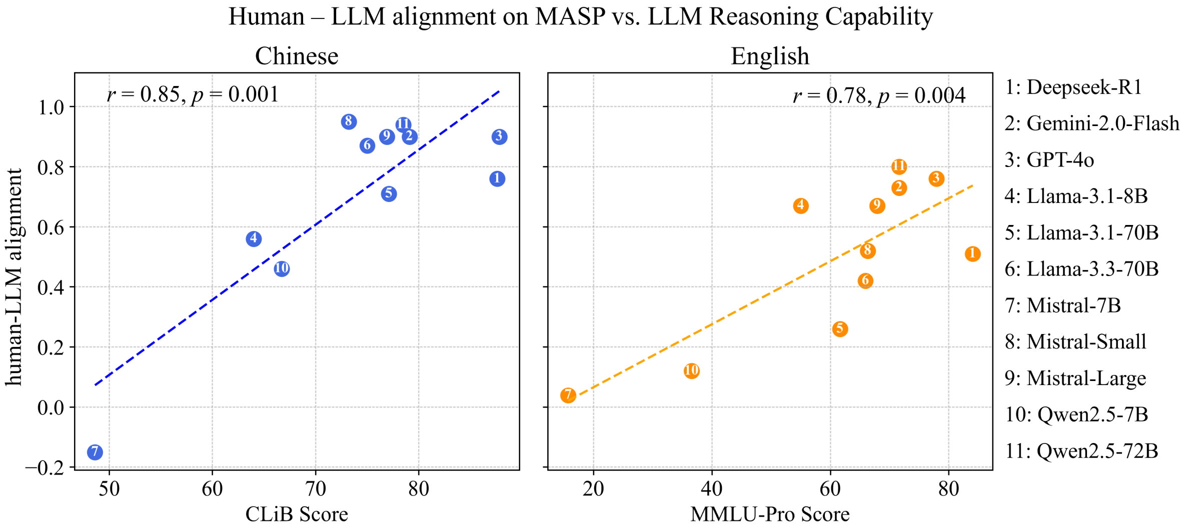


Figure 3: Correlation between human-LLM alignment and the reasoning capability of LLMs. The reasoning capability is quantified using the benchmark scores in CLiB (for Chinese) and MMLU-pro (for English).

(GPT-4o, Gemini-2.0-Flash, DeepSeek-R1, Qwen-2.5-72B, and Mistral-Large) on the MASP dataset. The correlations ranged from 0.89 to 0.97, indicating that these models exhibit highly consistent scalar judgments across languages and conditions.

3.2.2 Alignment between Human and LLM Behavior

We further analyzed the alignment between human and LLM behavior on the *degree modification* of MASP. To quantify the alignment, we computed the Pearson correlation between the proportion of entailment from humans and LLMs (Table 2). GPT-4o and Gemini-2.0-Flash showed human-aligned behavior in most tests of *degree modification*, while the behavior of smaller LLMs (e.g., Qwen2.5-7B and Llama-3.1-8B) were not strongly correlated with human behavior (Table 2). To characterize the relationship between human-LLM alignment on MASP and LLMs performance, we further measured the correlation between human-LLM alignment (Pearson correlation that averaged across all tests of *degree modification*) and model performance on two representative benchmarks for language comprehension: MMMU-Pro (Yue et al., 2024) for English and CLiB (EasyLLM, 2025) for Chinese. LLMs with higher benchmark scores (i.e., better reasoning capabilities in general) tended to perform more similarly to humans on the MASP dataset (Figure 3).

Overall, these results suggested that LLMs generally captured human-like degree semantics when adjectives were combined with a single modifier. Besides, the alignment between human and LLM behavior in MASP was strongly associated with the LLMs' reasoning abilities, highlighting the importance of evaluating formal semantics as a benchmark for evaluating human-like language understanding.

3.2.3 Modification Strength of Negated Modifier

We further analyzed five representative LLMs behavior on the tests of *compositional negation*. For bare adjectives, negation induced scale reversal by selecting the complementary degree interval (Figure 4). The proportion of entailment declined at the upper end of the scale on the modifier negation test. This pattern suggested that LLMs interpreted the modifier negation as the complement set of all adjective phrases (e.g., complement set of *very tall*). In contrast, on the adjective negation test, the proportion of entailment decreased more steeply and earlier, where phrases like "*very not tall*" were more readily interpreted as directly entailing "*very short*" (see Appendix Figure 2). These results revealed that LLMs could correctly distinguish the adjective negation and the modifier negation, showing a clear understanding of the negated modifier.

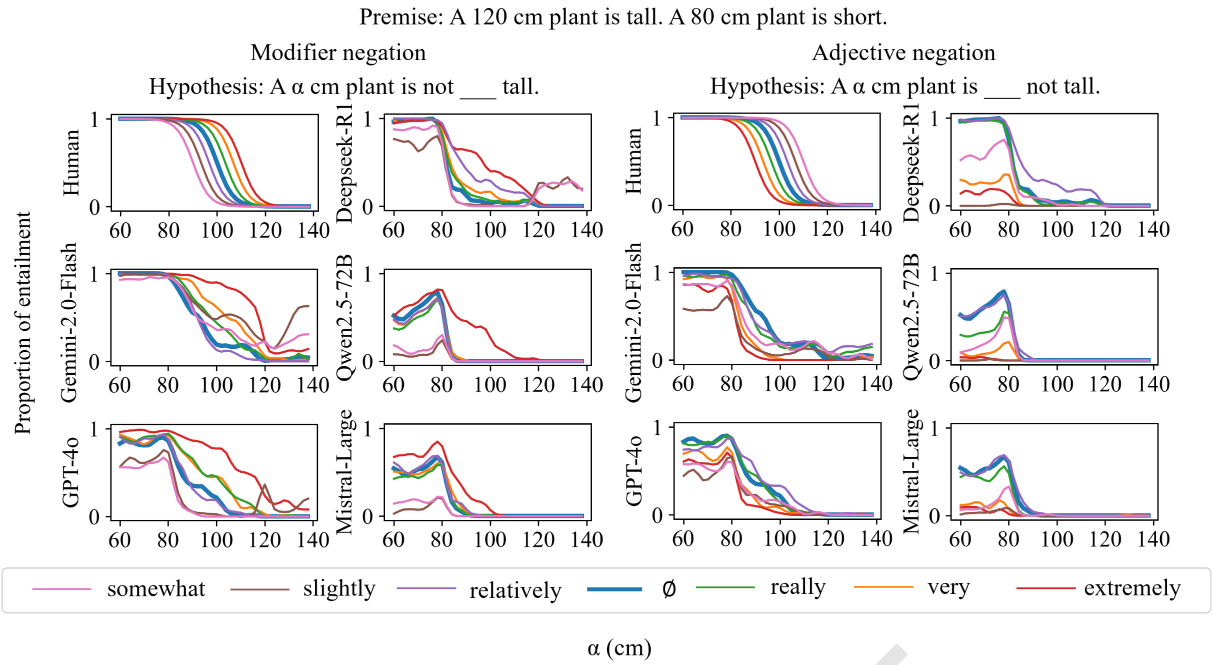


Figure 4: Theoretically expected human performance and actual model performance on the English tests of *compositional negation* with positive adjectives. The plots on the left illustrated cases of modifier negation, whereas those on the right represented adjective negation. The theoretically expected human performance is estimated based on the sigmoid function (Hersh and Caramazza, 1976), since we do not collect human data on compositional negation. See complete results in Appendix Figures 3, 4, 5 and 6.

4 Parameterizing Human and LLM Behavior

4.1 Experimental Setup

To systematically compare the human and LLM behavior, we fitted the proportion of entailment in the tests of *degree modification* using a few interpretable parameters. This approach was aligned with psychophysical frameworks for modeling semantic thresholds (Kruschke, 2014) and computational semantics methodologies for gradable expressions (Kennedy and McNally, 2005).

The proportion of entailment was fitted using a sigmoid function and a Gaussian function:

$$f(x) = \frac{A}{1 + e^{B(x-C)}} + amp * e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Where A, B, C, μ, σ, and amp are the parameters used for fitting. We defined three interpretable parameters to capture critical aspects of human and LLM behavior:

1. Polarity Confidence (A), which determined confidence in accepting hypotheses. Higher values of A indicated stronger confidence in entailment judgments (Lassiter, 2017).

2. Decision Threshold (C), which determined the position at which the proportion of entailment exceeded 0.5, indicating a transition from rejection to acceptance of entailment. Lower values of C implied tolerance for weakly intensified statements (e.g., *somewhat long*), while higher values of C indicate stricter thresholds, requiring modifiers with increased modification strength (Kennedy and McNally, 2005).

3. Judgment Precision (σ), which quantified the dispersion of responses. The lower values of σ indicate concentrated judgments (high precision), while the larger values of σ reflect the tolerance to ambiguity (Kruschke, 2014).

We employed non-linear least squares (NLLS) regression to fit the composite function to the proportion of entailment for LLMs and humans, minimizing the mean squared error (MSE). We performed 5-fold cross-validation and the averaged regression performance were reported in Appendix Table 1.

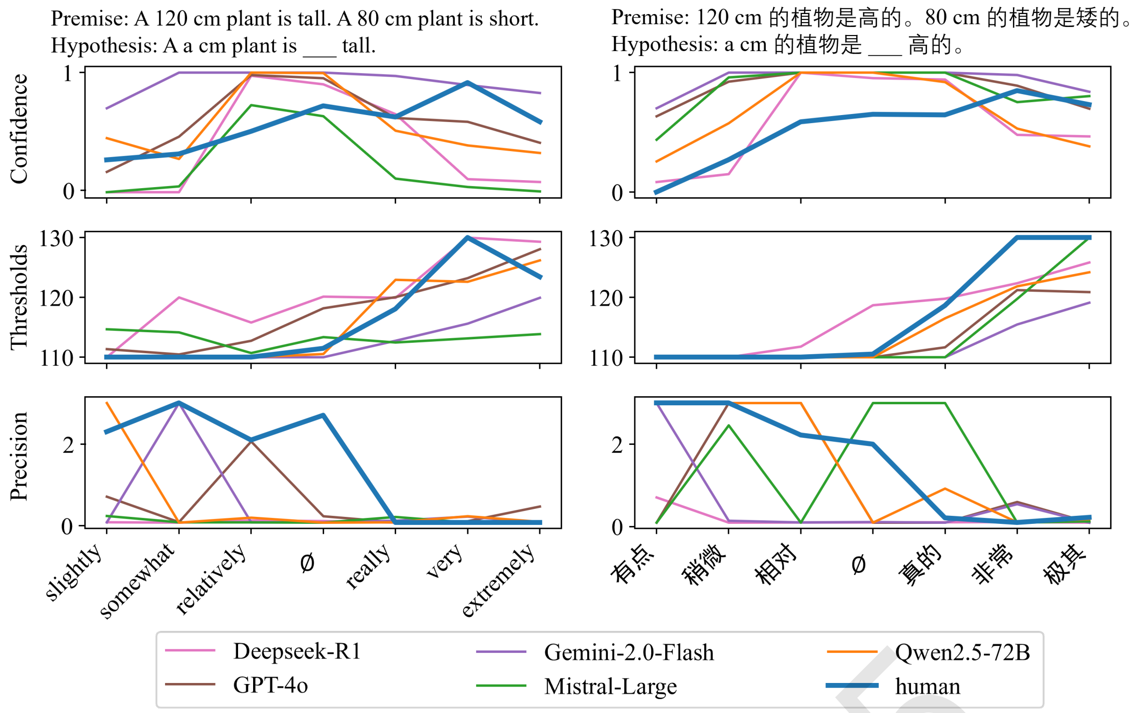


Figure 5: Polarity confidence, decision thresholds and judgment precision fitted from humans and five representative LLMs behavior.

4.2 Result

4.2.1 Polarity Confidence Varies with Modification Strength

As the modification strength of the modifier changes (e.g., from *somewhat long* to *very long*), the polarity confidence, representing the confidence in accepting hypotheses, did not increase monotonically in LLMs, whereas monotonicity was generally maintained in humans (see the top row of Figure 5). Instead, LLMs showed the highest polarity confidence when bare adjectives (∅) were presented, with confidence decreasing for both diminishers and intensifiers. This produced a trapezoidal pattern, suggesting that LLMs preferred to make more confident entailment decisions for single adjective rather than adjective phrases. In contrast, humans showed a more consistent upward trend, indicating growing confidence as the modification strength increased. This divergence suggested that while humans interpreted intensifiers as enhancing meaning, LLMs might treat them as softening the entailment signal.

4.2.2 Decision Thresholds Shift Along the Modification Strength

The decision thresholds, representing the transition from rejection to acceptance of entailment, systematically shifted along with modification strength for both humans and LLMs (see the middle row of Figure 5). For the intensifiers (e.g., *very*), the decision threshold shifted rightward (or leftward for negative adjectives). This result indicated that LLMs dynamically adjusted the internal threshold to categorize the premise-hypothesis pair as entailment or non-entailment, depending on different modifiers.

While both humans and LLMs exhibited the trend, humans tended to change more abruptly, completing the transition within just one or two neighboring modifiers, whereas LLMs showed a more gradual, smoothed adjustment. This suggested that humans recalibrated category boundaries more discretely, while LLMs distributed the adjustments more continuously.

4.2.3 Judgment Precision Increases for Extreme Modifiers

The judgment precision, which reflected how sharply the LLMs or humans distinguished entailment from non-entailment, did not consistently decrease with the strength of the modifier across LLMs (see the bottom row of Figure 5). While humans (blue line) showed a relatively clear trend toward increased precision for modifiers (e.g., from *slightly long* to *extremely long*), indicating that entailment decisions became more concentrated, most LLMs displayed non-monotonic patterns.

This contrast suggested that humans recalibrated with intensifiers in a more categorical and decisive fashion compared to diminishers, while LLMs might lack consistent internal scaling, leading to more diverse judgments. Altogether, these results revealed that despite showing graded sensitivity to degree modification, LLMs encoded the degree semantic information differently from humans.

5 Related Work

Understanding how LLMs process language is essential for model interpretability and the construction of more reliable language models. Despite the impressive performance of LLMs like GPT-4 on general-purpose benchmarks, a large number of studies suggested that these models remain fragile in domains requiring fine-grained semantic reasoning. Liu et al. (2023a) demonstrated that current LLMs have difficulties in modeling ambiguity, including vague modifiers and context-dependent interpretations. Relatedly, Lin et al. (2024) observed that even advanced models display inconsistent patterns in plan reasoning tasks. Our study contributes to this body of literature in theoretically informative ways. The MASP dataset focuses on the evaluation of scalar expressions, including gradable adjectives and adjective phrases. The MASP dataset is grounded in the formal semantics of adjectives (specifically degree semantics), leveraging its theoretical constructs to probe LLMs' internal representations. Parallel work within degree semantics has examined gradable adjective properties: Soler and Apidianaki (2021) demonstrated that diagnostic classifiers can predict adjective modification intensity from BERT's word embeddings. Liu et al. (2023b) revealed that language models encode the degree semantics of bare adjectives but fail to employ it. Both of these studies aim at the bare adjectives, while ignoring the compositionality between adjectives and modifiers, which is the key aspect of the adjective interpretation. Our study builds on this growing body of research by offering a diagnostic dataset for compositional semantics between adjectives and modifiers, going beyond prior work that typically investigates isolated adjective-modifier pairs.

In addition, the MASP dataset provides an approach to quantify the cross-linguistic interpretation of adjective phrases, i.e., Chinese and English. Though most LLMs have multilingual capabilities, most diagnostic datasets and benchmarks focus on model performance in English (Wang et al., 2019; Rajpurkar et al., 2016), which may lead to square-one bias for model evaluation (Ruder et al., 2022). Crucially, scalar reasoning should account for cross-linguistic generalization, as degree semantics and the effect of modifiers may differ across typologically diverse languages (Beck et al., 2021). Yet few probing datasets currently evaluate such behavior in multiple languages. Our work addresses this gap by constructing a multilingual dataset to analyze the cross-linguistic difference in degree semantic information encoded by LLMs. Therefore, we offer a theoretically motivated approach to evaluate LLMs' ability to encode the degree semantic information of adjectives and adjective phrases, and extend previous studies into a multilingual, compositional, and theoretically grounded setting.

6 Conclusion

In conclusion, building upon adjective degree semantics, this study constructs the MASP dataset to evaluate LLMs' understanding of adjectives and modifiers. By integrating scalar modifiers and negation modifier into the degree estimation task, we uncover systematic behavior through interpretable metrics: polarity confidence, decision thresholds, and judgment precision. While LLMs generally encode degree semantic information, LLMs' understanding of adjective phrases diverges from human patterns in key areas, particularly for relatively neutral modifiers. Our study offers a compact, scalable tool for probing degree semantics and advancing human-aligned, cross-linguistic language understanding.

While MASP relies on synthetic templates, this allows for controlled and cross-linguistically comparable evaluation. We acknowledge its limitations in naturalistic contexts and leave this for future work. Critically, the observed divergences point to a fundamental challenge for current LLMs: their reported deficiency in robust compositional reasoning, especially when processing nuanced interactions between modifiers and adjectives. To overcome this limitation, future work should prioritize brain-inspired computational models implementing sequence chunking through neural encoding of ordinal positions (Ding, 2025), alongside recent advances in: (1) Topological neural manifolds (Deb et al., 2025) enforcing brain-like spatial organization for compositional binding, (2) Dynamic composition attention (Xiao et al., 2024) closing the modifier-adjective integration gap.

Acknowledgements

The authors would like to thank the anonymous reviewers for their helpful suggestions and comments. This research is supported by STI2030-Major Project 2021ZD0204100 (2021ZD0204105).

References

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, Lisbon, Portugal. Association for Computational Linguistics.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Laura Ruis, Akbir Khan, Stella Biderman, Sara Hooker, Tim Rocktäschel, and Edward Grefenstette. 2023. The goldilocks of pragmatic understanding: Fine-tuning strategy matters for implicature resolution by LLMs. In *Proceedings of the 37th International Conference on Neural Information Processing Systems (NeurIPS 2023)*, article 913, New Orleans, LA, USA. Curran Associates Inc., Red Hook, NY, USA.

Christopher Kennedy. 2007. Vagueness and grammar: The semantics of relative and absolute gradable adjectives. *Linguistics and Philosophy*, 30:1–45. https://doi.org/10.1007/s10988-006-9008-0.

Daniel Lassiter. 2017. *Graded Modality: Qualitative and Quantitative Perspectives*. Oxford University Press, United States. https://doi.org/10.1093/oso/9780198701347.001.0001.

Alisa Liu, Zhaofeng Wu, Julian Michael, Alane Suhr, Peter West, Alexander Koller, Swabha Swayamdipta, Noah Smith, and Yejin Choi. 2023a. We‘re afraid language models aren‘t modeling ambiguity. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 790–807, Singapore. Association for Computational Linguistics. https://aclanthology.org/2023.emnlp-main.51/.

Fangru Lin, Emanuele La Malfa, Valentin Hofmann, Elle Michelle Yang, Anthony Cohn, and Janet B. Pierrehumbert. 2024. Graph-enhanced large language models in asynchronous plan reasoning. *arXiv preprint arXiv:2402.02805*. https://arxiv.org/abs/2402.02805.

Koustuv Sinha, Shagun Sodhani, Jin Dong, Joelle Pineau, and William L. Hamilton. 2019. CLUTRR: A diagnostic benchmark for inductive reasoning from text. 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 4506–4515, Hong Kong, China. Association for Computational Linguistics. https://aclanthology.org/D19-1458/.

R. Thomas 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, Florence, Italy. Association for Computational Linguistics. https://aclanthology.org/P19-1334/.

Wei Liu, Ming Xiang, and Nai Ding. 2023b. Adjective Scale Probe: Can Language Models Encode Formal Semantics Information? In *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(11):13282–13290. https://doi.org/10.1609/aaai.v37i11.26559.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A New Benchmark for Natural Language Understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics. https://aclanthology.org/2020.acl-main.441/.

Oyvind Tafjord, Bhavana Dalvi Mishra, and Peter Clark. 2022. Entailer: Answering Questions with Faithful and Truthful Chains of Reasoning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2078–2093, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. https://aclanthology.org/2022.emnlp-main.134/.

Carita Paradis. 2008. Configurations, construals and change: Expressions of DEGREE. *English Language and Linguistics*, 12(2):317–343. Cambridge University Press.

Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. London: Pearson Longman.

Laria Reynolds and Kyle McDonell. 2021. Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems (CHI EA '21)*, Article 314, 7 pages. Yokohama, Japan. Association for Computing Machinery. https://doi.org/10.1145/3411763.3451760.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems*, 33:1877–1901. https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*, 2407.21783. https://doi.org/10.48550/arXiv.2407.21783.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, William El Sayed. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2310.06825. https://doi.org/10.48550/arXiv.2310.06825.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024. Qwen2 Technical Report. *arXiv preprint arXiv:2407.10671*, 2407.10671. https://doi.org/10.48550/arXiv.2407.10671.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, et al. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *arXiv preprint arXiv:2501.12948*, 2501.12948. https://doi.org/10.48550/arXiv.2501.12948.

Google DeepMind. 2025. Gemini 2.0. https://deepmind.google/technologies/gemini.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, et al. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*. https://doi.org/10.48550/arXiv.2303.08774.

Stephanie Solt. 2019. Comparison to Arbitrary Standards. In *Proceedings of Sinn und Bedeutung*, 16(2):557–570. https://ojs.ub.uni-konstanz.de/sub/index.php/sub/article/view/447.

Christopher Kennedy and Louise McNally. 2005. Scale Structure, Degree Modification, and the Semantics of Gradable Predicates. *Language*, 81(2):345–381. https://dx.doi.org/10.1353/lan.2005.0071.

Harry M. Hersh and Alfonso Caramazza. 1976. A fuzzy set approach to modifiers and vagueness in natural language. *Journal of Experimental Psychology: General*, 105(3):254–276. https://doi.org/10.1037/0096-3445.105.3.254.

EasyLLM. 2025. Benchmarking of Open LLMs. https://easyllm.site/static/benchmarking.html.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. 2024. MMMU-Pro: A More Robust Multi-discipline Multimodal Understanding Benchmark. *arXiv preprint arXiv:2409.02813*. https://doi.org/10.48550/arXiv.2409.02813.

John Kruschke. 2014. Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan. *Academic Press*.

Aina Gari Soler and Marianna Apidianaki. 2021. Let's play mono-poly: BERT can reveal words' polysemy level and partitionability into senses. *Transactions of the Association for Computational Linguistics*, 9:825–844. https://aclanthology.org/2021.tacl-1.50/.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2019/file/4496bf24afe7fab6f046bf4923da8de6-Paper.pdf.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics. https://aclanthology.org/D16-1264/.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2022. Square One Bias in NLP: Towards a Multi-Dimensional Exploration of the Research Manifold. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2340–2354, Dublin, Ireland. Association for Computational Linguistics. https://aclanthology.org/2022.findings-acl.184/.

Sigrid Beck, Sveta Krasikova, Daniel Fleischer, Remus Gergel, Stefan Hofstetter, Christiane Savelsberg, John Vanderelst, and Elisabeth Villalta. 2021. Crosslinguistic variation in comparison constructions. *Linguistic Variation Yearbook*, 9(1):1–29. https://doi.org/10.1075/livy.9.01bec.

Nai Ding. 2025. Sequence chunking through neural encoding of ordinal positions. *Trends in Cognitive Sciences*. Elsevier.

Mayukh Deb, Mainak Deb, and N. Apurva Ratan Murty. 2025. TopoNets: High Performing Vision and Language Models with Brain-Like Topography. *arXiv preprint arXiv:2501.16396*. https://arxiv.org/abs/2501.16396.

Da Xiao, Qingye Meng, Shengping Li, and Xingyuan Yuan. 2024. Improving transformers with dynamically composable multi-head attention. In *Proceedings of the 41st International Conference on Machine Learning (ICML'24)*, Vienna, Austria, pages 2231–2249. JMLR.org. https://github.com/Caiyun-AI/DCFormer.

Lauscher, Anne, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From Zero to Hero: On the Limitations of Zero-Shot Language Transfer with Multilingual Transformers. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 4483–4499. Association for Computational Linguistics. https://aclanthology.org/2020.emnlp-main.363/.

Appendix

Appendix A: LLMs models experiment details

We adopted a zero-shot prompting setup for all LLM models, constructing inputs as natural language inference (NLI) tasks (Wang et al., 2019). Each model was instructed to determine entailment between a premise and hypothesis without in-context examples, ensuring reliance on pre-trained representations rather than task-specific adaptation (Brown et al., 2020). The prompts were structured as follows.

English Prompt: *Premise: {premise}, Hypothesis: {hypothesis}. Determine whether the premise entails the hypothesis. Respond with ’entailment’ or ’not entailment’.*

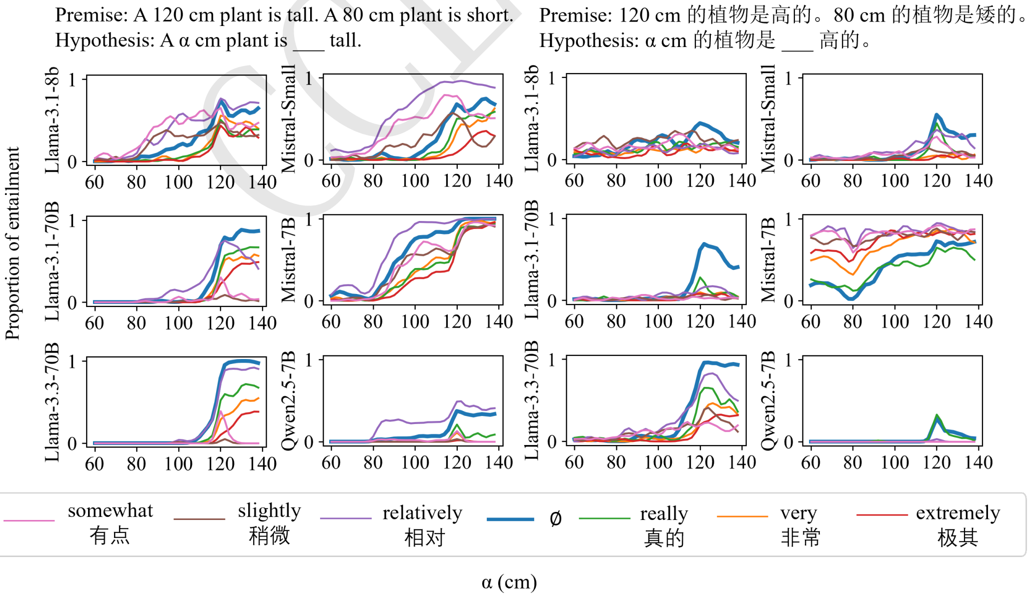
Chinese Prompt: 前提：{*premise*}，假设：{*hypothesis*}，判断前提是否蕴含假设。回答 ’蕴含’ 或 ’非蕴含’。 This design follows methodologies for probing cross-lingual semantic consistency in LLMs (Lauscher et al., 2020) while minimizing language-specific prompt engineering (Reynolds and McDonell, 2021).

Appendix B: Human annotation for the degree estimation task

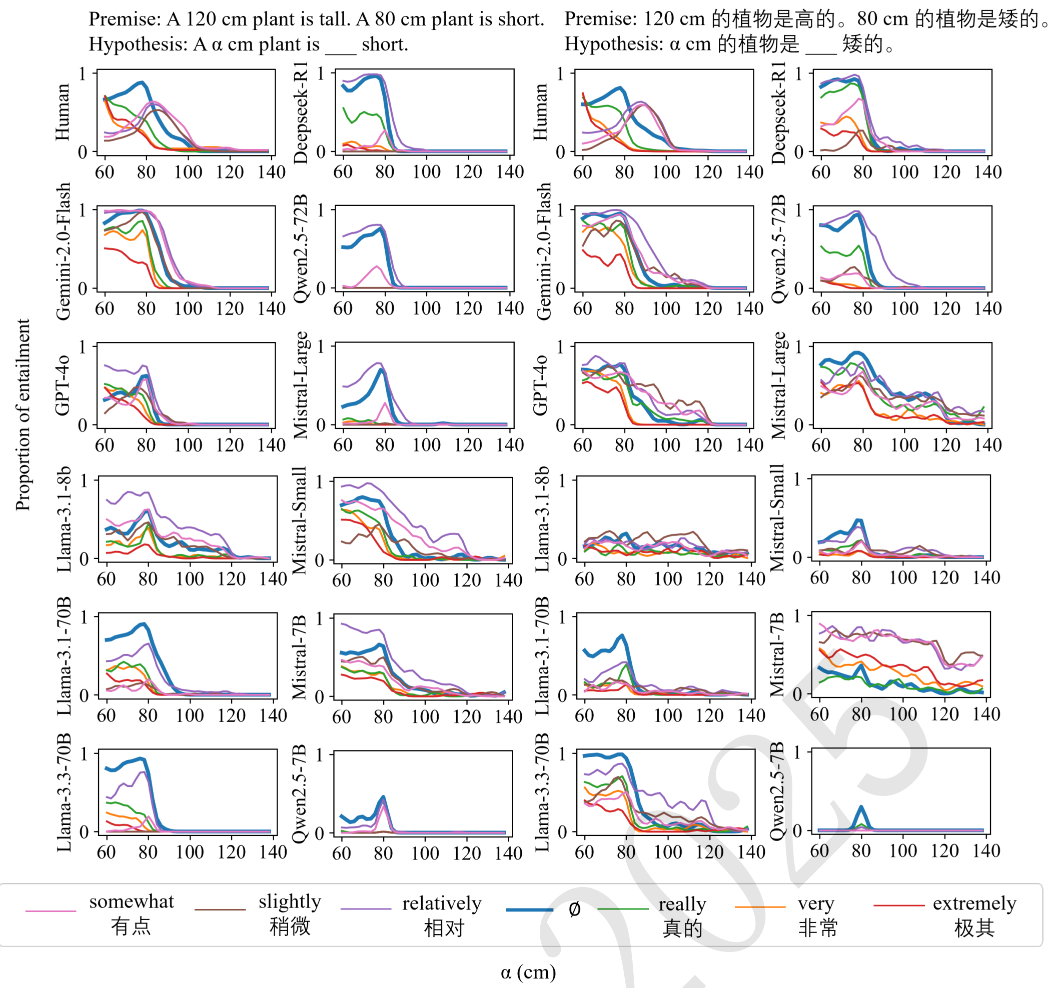
The human annotation for degree estimation task was designed in a structured numerical reasoning format to assess degree semantic interpretations under scalar modifiers. The English and Chinese version of the task followed the template as shown below:

English version:
If a 120 cm plant is tall, and a 80 cm plant is short,
Then,
A [modifier] *tall plant is between ____ cm and ____ cm.*
Chinese version:
如果 *120 cm* 的植物是高的。 *80 cm* 的植物是矮的。
那么，
____ *cm* 到____ *cm* 的植物是 [modifier] 高的。
modifier ∈ {∅, *very*, *really*, *extremely*, *relatively*, *slightly*, *somewhat*}

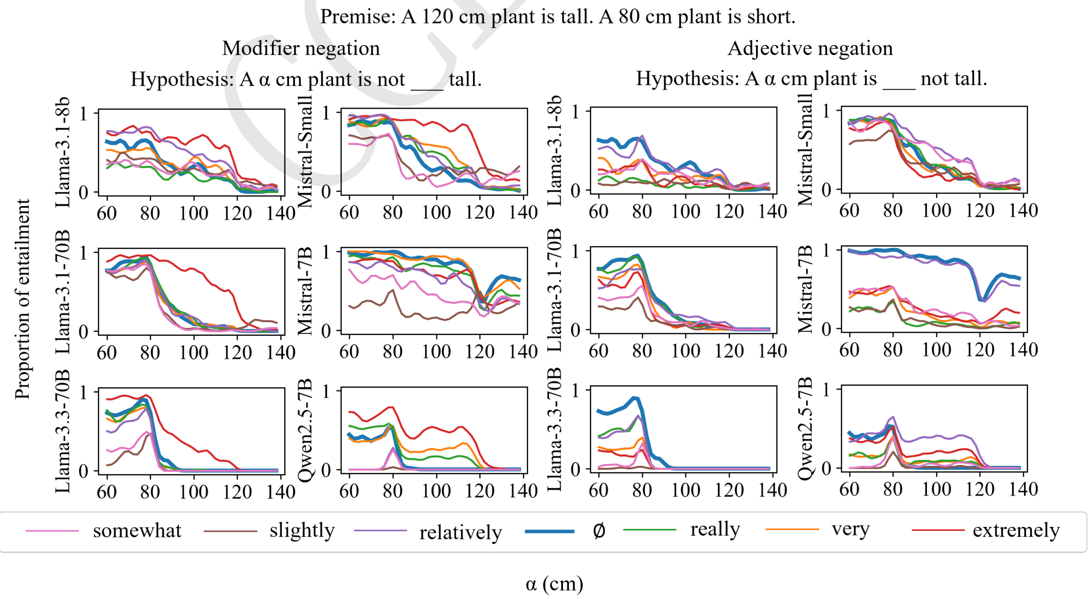
Appendix C: Human and model performance on both English and Chinese cross tests



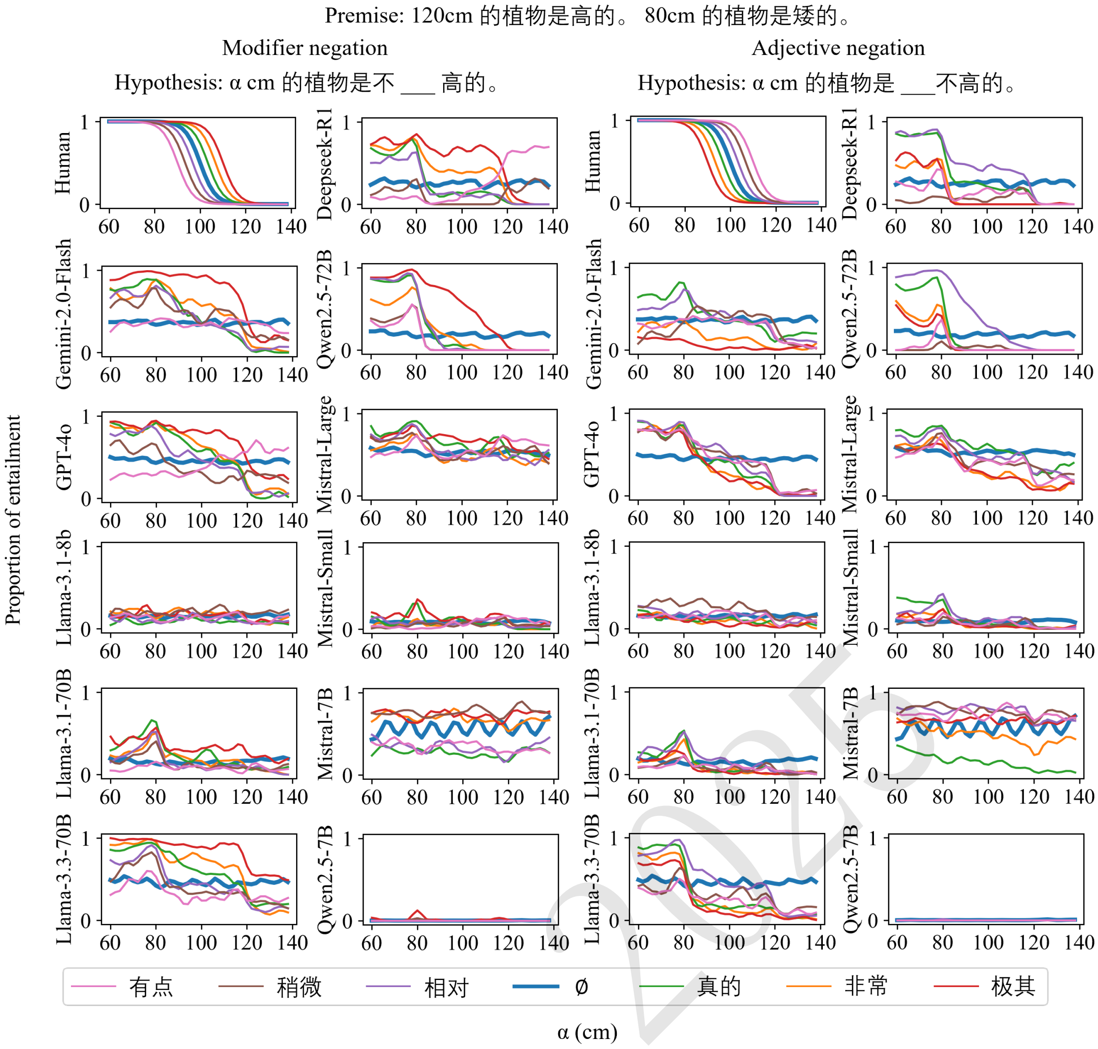
Appendix Figure 1: Other model performance on both Chinese and English tests of *degree modification* with positive adjectives.



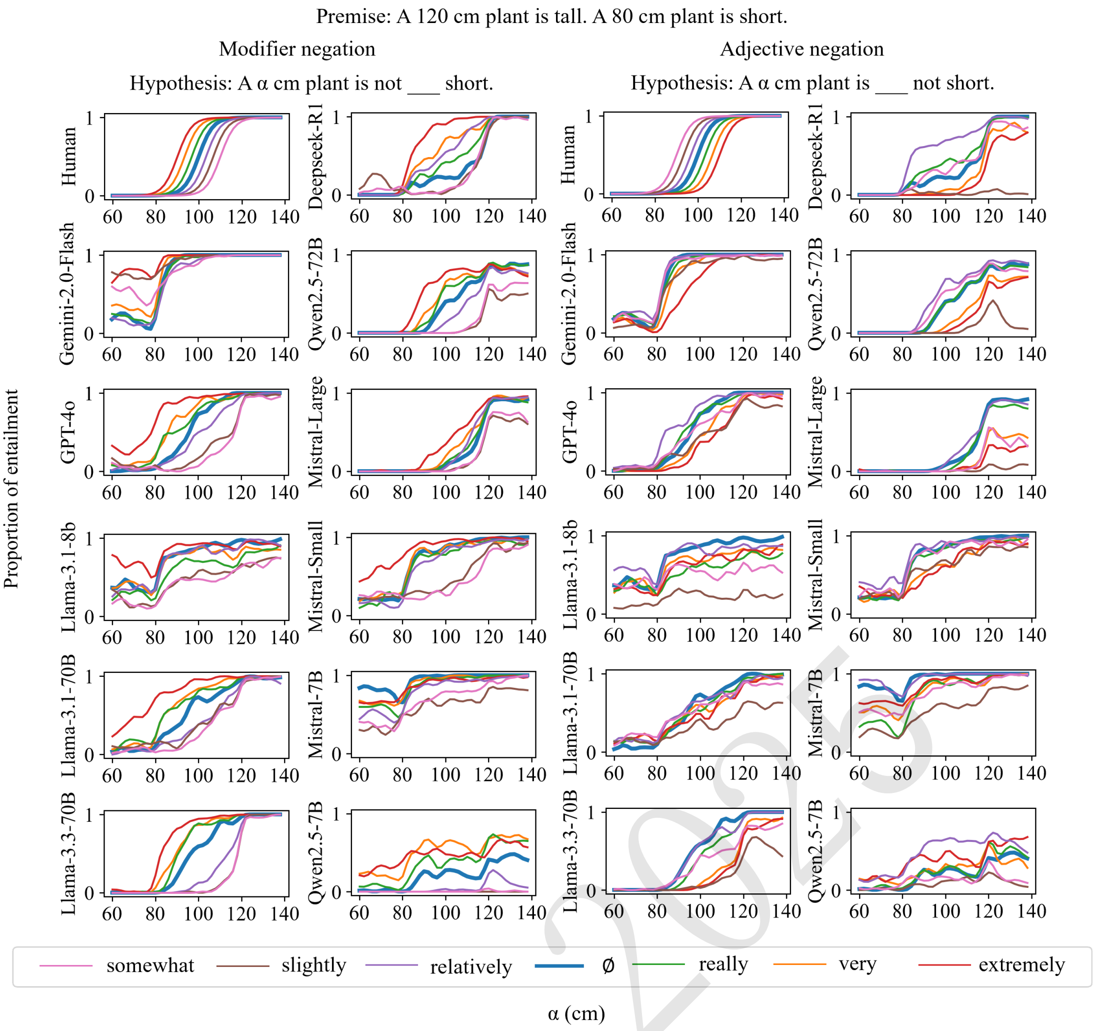
Appendix Figure 2: Human and model performance on both Chinese and English tests of *degree modification* with negative adjectives.



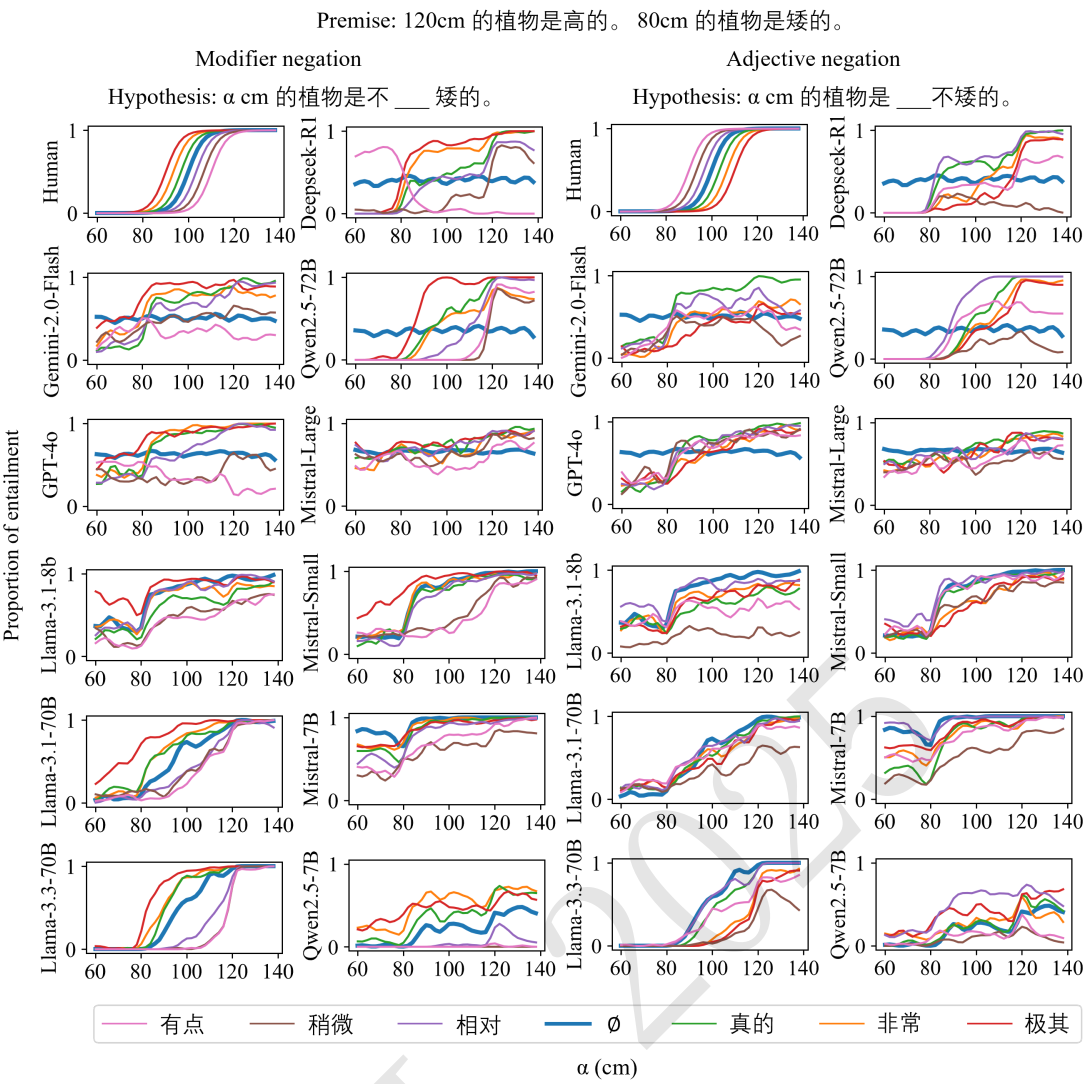
Appendix Figure 3: Other model performance on the English tests of *compositional negation* with positive adjectives.



Appendix Figure 4: Theoretically expected human performance and actual model performance on the Chinese tests of *compositional negation* with positive adjectives. The theoretically expected human performance is estimated based on the sigmoid function (Hersh and Caramazza, 1976), since we do not collect human data on compositional negation.



Appendix Figure 5: Theoretically expected human performance and actual model performance on the English tests of *compositional negation* with negative adjectives. The theoretically expected human performance is estimated based on the sigmoid function (Hersh and Caramazza, 1976), since we do not collect human data on compositional negation.



Appendix Figure 6: Theoretically expected human performance and actual model performance on the Chinese tests of *compositional negation* with negative adjectives. The theoretically expected human performance is estimated based on the sigmoid function (Hersh and Caramazza, 1976), since we do not collect human data on compositional negation.

Appendix D The R^2 of the curve fitting for the entailment responses of the LLM models on test fold

Model	Chinese							English						
	∅	非常	真的	极其	相对	稍微	有点	∅	very	really	extremely	relatively	slightly	somewhat
Human	0.85	0.79	0.78	0.87	0.8	0.75	0.87	0.79	0.79	0.9	0.78	0.83	0.84	0.8
Deepseek-R1	0.89	0.86	0.86	0.9	0.89	0.82	0.83	0.87	0.9	0.88	0.8	0.75	0.82	0.83
GPT-4o	0.81	0.85	0.8	0.86	0.91	0.86	0.78	0.87	0.91	0.81	0.87	0.86	0.87	0.86
Gemini-2.0	0.87	0.8	0.74	0.83	0.71	0.87	0.71	0.91	0.89	0.8	0.9	0.71	0.79	0.83
Mistral-Large	0.87	0.87	0.76	0.75	0.86	0.83	0.85	0.83	0.79	0.83	0.86	0.78	0.87	0.91
Qwen2.5-72B	0.91	0.81	0.73	0.81	0.87	0.71	0.91	0.86	0.8	0.86	0.8	0.88	0.83	0.82

Appendix table 1: R^2 of the curve fitting on both Chinese and English tests of *degree modification*.