CrosSing: Cross-Scale Reasoning Evaluation on LLMs against Humans

Qi Han[*]
Cornell University
Linguistics Department
qh246@cornell.edu

Yifan Wu[*]
Cornell University
Linguistics Department
yw2578@cornell.edu

Marten van Schijndel
Cornell University
Linguistics Department
mv443@cornell.edu

Abstract

While many studies have shown LLMs perform well in various reasoning tasks, few have examined their capacity on semantic reasoning tasks. As LLMs reason with language, it is crucial to understand how well they grasp and use the underlying scalar relationships in language. In this study, we introduced a new dataset **CrosSing** (Cross-Scale reasoning), providing a human baseline against which to evaluate LLMs' ability to reason across lexical scales in gradable adjectives. We further probed how their understanding is influenced by overinformative contexts. We evaluated ten high-performing LLMs and found that some outperformed humans when no extra information was provided, but that LLM performance declined in certain overinformative contexts while human performance improved significantly. This contrast reveals a fundamental difference between recent LLMs and humans in understanding adjectives' scalar relationships and how such understanding behaves in overinformative contexts.

1 Introduction

Opposing scales are common in real life use cases. Consider the following example shown in Figure 1, a chocolate seller categorizes his chocolates into four specific tiers: free, cheap, expensive, and luxurious, and he asks an LLM-based chatbot for pricing strategies based on the tiers. To give a reasonable response, the chatbot needs to rank the four tiers on a unified scale to map their relative orderings to appropriate price values.

Tasks like this chocolate-pricing example rely on the ability to map and order adjectives from multiple scales onto a single shared dimension. This can occur with abstract properties such as affordability and concrete scalar values like prices. Such an ordering usually involves an understanding of scalar implicatures (SI; Grice, 1975; Horn, 1989; Levinson, 2000, a.o.), where listeners draw an inference that a speaker's use of a weaker expression implies the negation of a stronger alternative (e.g., *expensive* → "*expensive but not luxurious*"). While human SI reasoning is well-explored (e.g., Noveck, 2001; Papafragou and Musolino, 2003; Bott and Noveck, 2004; Pouscoulous et al., 2007; Katsos and Bishop, 2011; Frank and Goodman, 2012; Goodman and Stuhlmüller, 2013; Degen, 2015; Degen and Tanenhaus, 2015, 2016; Goodman and Frank, 2016; Skordos and Papafragou, 2016; Gotzner et al., 2018; Kampa et al., 2025), less is known about whether large language models (LLMs) align with humans in representing and reasoning with adjectives across opposing scales.

Figure 1: Example of cross-scale reasoning. Picture generated by Nano Banana Pro.

Recent computational work has shown that language models can induce scalar structures and encode SIs in a parallel fashion to human pragmatic reasoning (Kim and de Marneffe, 2013; Schuster et al., 2020; Soler and Apidianaki, 2020; Hu et al., 2023; Nizamani et al., 2024, a.o.). However, previous studies have focused on single-scale SIs. The present study contributes to this line of work by proposing a new task that asks human participants and LLMs to reason **across** scales in different information contexts.

We tested human and LLM **(i)** general understanding of cross-scale adjectival relations (Sec-

[*]Equal contribution.

Proceedings of the Society for Computation in Linguistics (SCiL) 2026, pages 379–407

tion 5), and **(ii)** ability to integrate task information (Section 6.1) or semantic information (Section 6.2) in cross-scale reasoning. As LLMs are increasingly trained to perform human tasks, it is crucial to have a parallel LLM-human comparison. LLMs encounter gradable adjectives frequently and are expected to reason about their underlying scales to map words into a coherent meaning space. If model ordering of these fundamental concepts differs from humans, then problems which depend on the accurate usage of these adjectives will naturally emerge. Our results show that while models can order cross-scale adjectives accurately in isolation, their representations are easily disturbed by extra information that caused no harm for, and in some cases aided, human reasoning.

2 Related Work

Gradable adjectives instantiate context-sensitive scales that yield scalar inferences (Beltrama and Xiang, 2013; Ronai and Xiang, 2022, a.o.). Experimental work has shown significant scalar diversity across scales, such that some scales prompt near-categorical implicatures, whereas other scales elicit low SI rates (van Tiel et al., 2016; van Tiel and Schaeken, 2017; van Tiel and Pankratz, 2021; Ronai and Xiang, 2023; Aparicio and Ronai, 2024, 2025). Building on these insights, recent work has explored how LLMs represent and reason over scalar structure.

Researchers have operationalized LM implicature modeling in various ways. Kim and de Marneffe (2013) introduced a distributional method that interpolates between word-embedding vectors to induce adjectival scales. Following Kim and de Marneffe (2013), subsequent methods using neural embeddings aimed to recover known scalar relations from distributional representations, achieving similar results in predicting human-like scalar rankings. For instance, Soler and Apidianaki (2020) leveraged BERT (Devlin et al., 2019) to rank gradable adjectives by intensity, and van Tiel et al. (2019) confirmed the expected scale orderings of adjectives against corpus data. Schuster et al. (2020) trained a neural classifier on human judgment data to predict SIs and showed that some linguistic features can suffice for accurate SI classification. Fine-tuned T5 models' probability estimates and expectation scores for omitted strong alternative adjectives closely mirror human SI variability (Hu et al., 2022, 2023). More recently, SIGA, a large

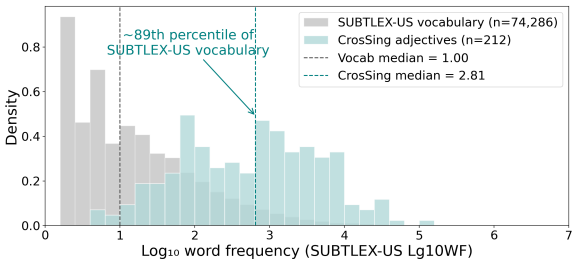

Figure 2: CrosSing Items Frequency Analysis: The median frequency of adjectives in CrosSing is at the 89th percentile of SUBTLX-US vocabulary.

natural language inference (NLI) benchmark for SIs among gradable adjectives by Nizamani et al. (2024), reported that fine-tuning improves detection accuracy of SIs while fine-grained contextual reasoning with gradable adjectives remains challenging. Similarly, broader NLI evaluation benchmarks such as the IMPRES dataset (Jeretic et al., 2020) and the Pragmatics Understanding Benchmark (PUB; Sravanthi et al., 2024) spanning over multiple pragmatic phenomena, demonstrate that fine-tuning LLMs on pragmatic examples markedly boosts performance. Nevertheless, modern LLMs still struggle to integrate subtle context cues across a wide range of pragmatic inferences, not limited to SI (Parrish et al., 2021; Qiu et al., 2023; Ruis et al., 2023; Cho and Kim, 2024, a.o.).

However, these existing studies have focused on how LMs encode and retrieve information within individual lexical scales, whereas no work has systematically examined their ability to pragmatically reason across multiple scales for scale composition.

3 CrosSing

To test how humans and models compose scales from gradable adjectives, we designed the dataset **CrosSing** (Cross-Scale Reasoning Evaluation), where each item consists of four adjectives from two opposing SI scales (e.g., [*free, cheap, expensive, luxurious*]). A total of 109 items were curated with 116 unique SI scales. Among those, 59 single-scale lexical pairs are from Aparicio and Ronai (2025), and another 8 pairs are from Nizamani et al. (2024), to which we added 49 pairs.[1]

The frequency of adjectives used in **CrosSing** ranges from the 28th to 99.9th percentile of the SUBTLEX-US dataset by Brysbaert et al. (2012), whose 74k vocabulary frequency is extracted from

[1]CrosSing: https://github.com/qihan-ling/CrosSing

cheap	free	expensive	luxurious
W_1	W_2	W_3	W_4
Two SI scales		$\langle W_1, W_2\rangle, \langle W_3, W_4\rangle$	
VALID RANKINGS		**4312, 2134**	
INVALID RANKINGS		22 permutations of 1–4	
COMMON WRONGS		1234, 4321, 2143, 3412	

Table 1: Example Item from Dataset and Possible Rankings. The numbers 1 and 3 refer to weak adjectives while 2 and 4 refer to strong ones in opposing SI scales 1-2 and 3-4. These rankings are order-invariant. Common Wrongs are top 4 error types from human experiments.

51 million words of American TV and movie subtitles. The median **CrosSing** adjective is more frequent than 89% of the SUBTLEX-US vocabulary (See Figure 2).

In each SI scale (e.g., [*expensive, luxurious*]), there is a strength comparison between the weak adjective (e.g., *expensive*) and the strong alternative (e.g., *luxurious*). Additionally, the two opposing scales share a unified scale (e.g., *'value'* for the example in Table 1). Items across these opposing single SI scales form antonymous pairs of varying intensity (e.g., [*cheap-expensive*, *free-luxurious*, *cheap-luxurious*, *free-expensive*]).

With four words in an item, there are 24 possible permutations to order them. When the goal is to rank them on the composed scale, there are only **two** satisfactory ranking results as illustrated in Table 1. For any item, the valid rankings must satisfy two conditions: (1) two words in the same SI scale and weak words of opposing SI scales are always direct **neighbors** to each other in the valid rankings (e.g., [*cheap-free, expensive-luxurious, cheap-expensive*]); (2) the weak **antonym** pair always occupies the middle positions (e.g., [*..., cheap , expensive...*]) and the strong pair sits on the two ends (e.g., [*free, ..., ..., luxurious*]). These two rankings for each item are treated as ground truths in the experiments below.[2]

4 Models

We selected ten models from five model families to test via web interfaces: *ChatGPT 4o* (Achiam et al., 2023), *ChatGPT o4-mini-high* (OpenAI, 2025), *Claude 3.7 Sonnet* (Anthropic, 2025), *Gemini 2.0 Flash*, *Gemini 2.5 Pro* (Comanici et al., 2025), *Qwen2.5 Max*, *Qwen3-30B*, *Qwen3-235B* (Yang et al., 2025), *DeepSeek R1* (Guo et al., 2025), *DeepSeek Prover V2* (Ren et al., 2025). These models include both open-sourced and closed-sourced language models. For every zero-shot prompt, we tested it on a separate chat to avoid contamination from other prompts. For ChatGPT models that are sensitive across conversations, we used its temporary chat functionality to disable it from using separate conversations as contexts.

[2]Equivalently, one may describe our analyses as reversal invariant.

5 Just Ranking

We zero-shot evaluated the chosen LLMs on the **CrosSing** task using three ranking conditions that we concurrently used to evaluate human scalar composition in a separate manuscript. Human participants completed a spatial drag-and-drop task, adapted from Starr et al. (2025), requiring them to rank four randomized adjectives along a continuous line, see Figure 3. Both humans and LLMs were evaluated under these three conditions: (1) ranking with no additional information (Just Ranking), (2) ranking with extra task information, and (3) ranking with extra semantic information.

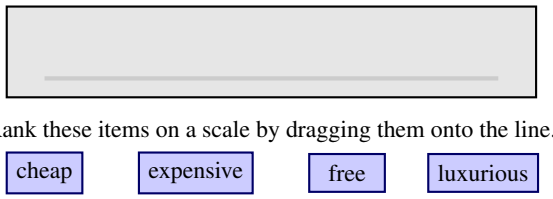

Figure 3: Human study item example, whose valid ordering is demonstrated in Table 1.

5.1 Human Baseline

Our human study established a performance baseline of 76% accuracy at this task, reflecting the extent to which participant responses aligned with the valid rankings, suggesting that human reasoners can map opposing scalar pairs onto a single semantic continuum. We noticed a significant positive correlation between item frequency and accuracy ($\rho = 0.23$ with $p < 0.01$), indicating that less familiar words are more difficult for humans to order. Among the possible 22 invalid rankings, 19 unique permutation types were present in human production. Four dominant wrong rankings (Table 1) simply concatenate the two opposite lexical scales (e.g., [cheap $\rightarrow$ free $\not\rightarrow$ expensive $\rightarrow$ luxurious] or [free $\rightarrow$ cheap $\not\rightarrow$ luxurious $\rightarrow$ expensive]). These dominant wrong rankings preserve the antonym relationships between the two

Model	$Acc\%$	U	ρ	$Acc_N\%$	U_N	ρ_N	$Acc_A\%$	U_A	ρ_A
Qwen3-235B	15.60	23	0.42	35.44↑	25	**0.79**	41.20↑	27	**0.78**
Qwen3-30B	65.75	14	**0.63**	28.81	25	**0.74**	32.19	25	**0.56**
Qwen2.5 Max	69.42	7	**0.56**	59.36	24	**0.79**	54.69	24	**0.63**
ChatGPT 4o	7.65	24	**0.45**	17.94↑	24	**0.76**	18.25↑	24	**0.82**
ChatGPT o4-mini-high	**94.50**	6	**0.8**	**87.67**	27	**0.81**	74.00	10	**0.84**
Gemini 2.0 Flash	**82.87**	9	**0.57**	18.86	24	**0.71**	49.80	24	**0.78**
Gemini 2.5 Pro	**94.8**	6	**0.8**	**95.82**↑	12	0.51	**80.65**	16	**0.79**
Claude 3.7	**78.90**	9	**0.7**	63.61	22	**0.82**	70.23	16	**0.76**
Deepseek-R1	**85.32**	12	0.5	54.57	20	**0.68**	75.03	25	**0.68**
Deepseek-Prover-v2	**88.69**	7	**0.6**	28.20	30	**0.76**	46.84	25	**0.68**
Human (baseline)	**76**	**21**	-	**81**	**19**	-	**76**	**20**	-

Table 2: All Experiments Results: Ten LLMs & human Acc (average accuracy), U (the number of unique ranking responses), and ρ (each model's response pattern's spearman correlation to the human baseline, bold values means $p < 0.01$, and values are highlighted if higher than 0.7), '$_N$' means the metric values are for Extra Task Information experiment, and '$_A$' refers to the Extra Semantic Information experiment.

SI scales and the weak-strong relationships within each subscale, but the composition of the two scales is flawed. We contrast these results with the LLMs results below.

5.2 LLMs Mis-/Super-Alignment

To avoid ordering effects of multi-turn conversations, we provided all 109 experimental items in one prompt (see prompts in Appendix 5). Three randomized orderings of items were run in separate conversations for each model. We found that six out of ten models outperformed humans on accuracy and had a substantially smaller range of wrong answers (Table 2). Models like ChatGPT o4-mini-high and Gemini 2.5 Pro achieved 94% accuracy, 18% higher than the human baseline with 15 fewer wrong answer types. The two models that produced similar numbers of unique ranking responses to humans (Qwen3-235B and ChatGPT4o) had the lowest accuracies (16% and 8%).

While higher-performing models like Gemini 2.5 Pro and ChatGPT o4-mini-high have a smaller range of answer rankings, they showed a high Spearman correlation to humans in terms of answer patterns ($\rho = 0.8$ with $p < 0.001$). Models like Qwen3-30B and Qwen2.5 Max with lower accuracies than humans exhibited a moderate correlation as well. The two worst-performing models (Qwen3-235B and ChatGPT4o) have the lowest correlation with human responses because they assign probability roughly evenly across the full range of rankings, in contrast to a skewed distribution over correct answer types which we see in humans and in other models.

Comparing models of the Qwen and ChatGPT families revealed that increasing model size does not ease the difficulty of the task. We found that smaller models exhibited more human-like response patterns and even had higher performance than humans (Qwen3-30B vs Qwen3-235B, and ChatGPT o4-mini-high vs ChatGPT 4o).

Given that models with better performance have a much smaller answer space than humans, we hypothesize that the training of models may have restricted their output space to the most common patterns that humans tend to produce, resulting in a peakier distribution over possible rankings than humans exhibit.

We observed that the error distribution for six out of ten LLMs have was significantly positively correlated with humans. Though ChatGPT4o has the lowest correlation of $\rho = 0.45$ to humans over all answer types, it shows moderate positive rank correlation with humans over only invalid answers ($\rho = 0.61$). That is, the errors that humans make more often, ChatGPT 4o also makes slightly more often (See Table 9 in Appendix G). In contrast to ChatGPT 4o, Gemini 2.0 Flash has a moderate correlation of $\rho = 0.57$ over all answer types, but its correlation to humans over only invalid answers is insignificant, implying that while it correctly produces more correct answers, it has different preferences over wrong rankings from humans. The correlation analysis indicates that models performing better than humans on this task do not always align with human preferences of possible ranking

Model	JR	R_N	R_A
Qwen3-235B	0.01	0.09	**0.34**
Qwen3-30B	**0.26**	0.13	**0.35**
Qwen2.5 Max	0.07	0.12	0.16
ChatGPT 4o	**-0.25**	**0.27**	0.23
o4-mini-high	**0.26**	0.08	**0.28**
Gemini 2.0 Flash	**0.31**	0.15	0.16
Gemini 2.5 Pro	**0.30**	0.10	0.13
Claude 3.7	0.22	0.04	**0.39**
Deepseek-R1	0.22	0.03	0.18
DS-Prover-v2	0.23	0.14	**0.31**

Table 3: Models Correlation of Item-wise Accuracy to Humans: JR(Just Ranking), R_N(Ranking with Extra Task Info), R_A(Ranking with Extra Semantic Info). Values with $p < 0.01$ significance are highlighted.

types and worse-performing models can still align with humans in their error distributions.

We further find that seven LLMs are positively correlated with humans in terms of item-wise accuracy. We note that having a positive correlation of ranking patterns does not entail a positive correlation of item-wise accuracy, as ChatGPT 4o has 0.61 ranking patterns correlation with $p < 0.01$ significance, but it has a negative Spearman correlation for item-wise accuracy. This means that while ChatGPT 4o makes similar ordering mistakes to humans, it performs worse on items humans find easy to rank and better on items humans find difficult.

While human results show a weak positive correlation between accuracy and word frequency ($\rho = 0.23$ with $p < 0.05$), none of the models exhibited behavior that was significantly correlated with word frequency (see Figure 6 in Appendix C).

6 Ranking with Extra Information

The super-human accuracy of many models in the Just-Ranking experiment might be because they have been explicitly trained to perform scalar inference. In this section, we tested the depth of their understanding by providing additional information which shouldn't alter their behavior. The motivation here is that reasoning across lexical scales require both understanding of the words (semantics of the words) and translating the representation onto a linear order (ordering as a task). With extra information about these two aspects of the task, we explore whether model behavior can be disrupted by these additions. At best, we expect that models could use the information to better encode the needed information. At worst, models can simply ignore the additional information.

6.1 Extra Task Information

To add extra task information to the Just-Ranking experiment, we appended additional information about the direct **neighbor** of two specific adjectives in the form of '*the word* [A] *and the word* [B] *should be placed next to each other*' (e.g., '*A is "cheap" and B is "expensive"* for ranking the list [free, cheap, expensive, luxurious]'). The neighboring information is always consistent with a valid ranking. To account for prompt formatting effects in LLMs, we used three prompts templates, where extra information can either be given once-for-all in a general manner, or repeated after each item with specific references to the words (see prompts in Appendix 5). Each prompt template was run with three random initializations, and we took the average accuracy as results reported below.

6.1.1 Enhanced Human Performance

When exposed to the extra task information, human participants achieved a mean accuracy of 81%, 5% higher than the Just Ranking baseline. The four dominant error rankings remained the same as those in the Just Ranking experiment. The mean correlation between word frequency and accuracy dropped to 0.03 from 0.23 in the Just Ranking experiment (See Figure 7 in Appendix C). Compared with results from Just Ranking, we observed that the lower frequency items' accuracies increase while the top most frequent items' accuracies dropped slightly (See Appendix G). This demonstrates that humans actively used the extra task information to help their ranking performances.

6.1.2 LLM Performance Drop

In contrast to the human improvement, seven out of ten models exhibited significantly degraded accuracy, ranging from -6% to -64% (Table 2).[3] Only ChatGPT o4-mini-high and Gemini 2.5Pro still surpassed human performance. Among the three improved models (Qwen3-235B, ChatGPT 4o, and Gemini 2.5 Pro), Qwen3-235B and ChatGPT 4o still underperformed with average accuracies of 35.44% and 17.94%.

[3]Some models produced incomplete rankings, resulting in more answer types than 24.

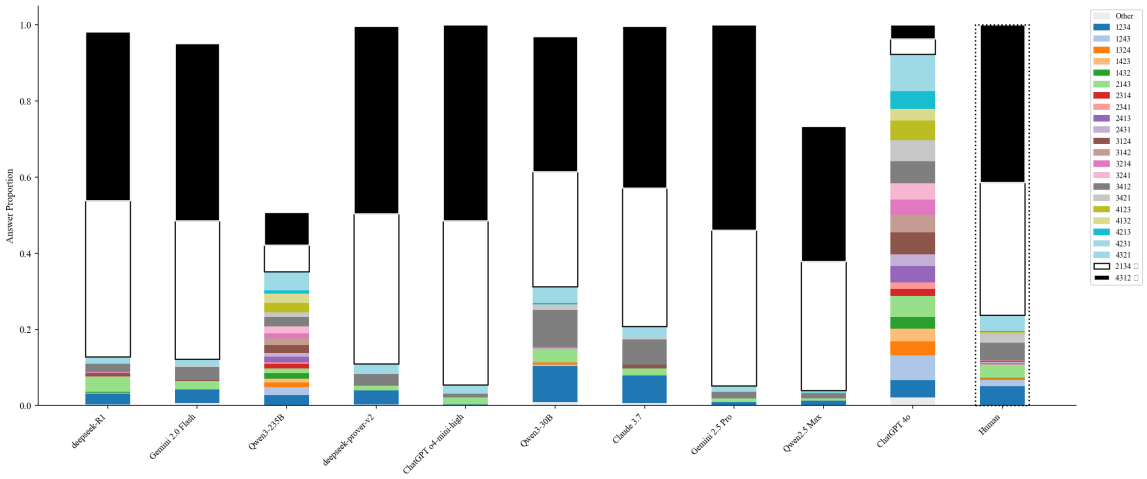

Figure 4: Just Ranking - Response Patterns of LLMs and Humans. Each color bar corresponds to a ranking type. For visual simplicity, low proportion rankings were grouped in the light grey 'other' bar. Black and white bars refer to two order-invariant valid ranking types. NA values (missing responses) are not included, so some models' answer proportions do not add to 1.

All models added at least six new invalid ranking responses over the Just-Ranking experiment. This trend underlines that the extra task information serves more as noise for the models than as guidance for the ranking task. As the extra task information provides partial information to check the correctness of the ranking, it should narrow the range of possible ranking patterns and increase the task accuracy at the same time. It remains a question about why some models (Gemini 2.5 Pro, Qwen3-235B, and ChatGPT 4o) had more kinds of invalid ranking responses but also better accuracies, indicating that, for those models, consideration of more potential rankings co-occurs with higher weights on the correct answers.

Looking closely at model answer distributions (Table 2), we found that all models except Gemini 2.5 Pro have significant strong positive correlation with human ranking patterns ($p < 0.001$), higher than their correlation to humans in the previous Just Ranking experiment. This pattern remained for correlations of error distributions as well (Table 9 in Appendix D). This suggests that the extra task information is noisy instead of informative for models, and when models got disrupted by such noise, they fell back to the data distribution learned from training and generated invalid rankings from the biased human distribution.

Though Gemini 2.5 Pro improved its accuracy by 1% with additional task information, its Spearman correlation of item-wise accuracy to humans decreased to 0.10 from 0.30 in the Just-Ranking experiment (Table 3). This drop in item-wise correlation score means that Gemini 2.5 Pro deviated from its previous agreement on items' difficulty levels to humans. In contrast, ChatGPT 4o's item correlation flipped from negative (-0.25) to positive (0.27), demonstrating a more aligned agreement on item difficulties to humans when given extra task information.

We also find that in this task both Gemini models (2.0 Flash and 2.5 Pro) show significant weak positive correlations between item frequency and accuracy (Figure 7 in Appendix C), while humans and other eight models do not. Gemini 2.5 Pro's correlation score between item frequency and accuracy increased from 0.13 in the Just Ranking experiment to 0.3 ($p < 0.01$) with additional task information, which is the opposite trend we see with humans from 0.23 ($p < 0.05$) in the Just Ranking experiment to 0.03 in current task. This implies that Gemini seems to be more influenced by word frequency than humans in generating rankings with extra task information.

6.1.3 Discussion

ChatGPT o4-mini-high and Gemini 2.5 Pro are the only two models with higher accuracies than humans in this task. Four out of six models that outperformed humans in the Just-Ranking experiment exhibited degraded accuracy to below the human baseline when presented with extra task information about how two words should be put next to each other. Such disruption of those models' behavior indicates that their understanding of

the relations between gradable adjectives is very brittle.

In addition to decreased performance, however, all models except Gemini 2.5 Pro exhibit similar ranking behavior to humans. One possible interpretation for this stronger ranking correlation is that models are initially overtrained or fine-tuned to make basic SI judgments but with additional task information, they fall back on the human data distribution learned during pre-training. However, we also find that the item-wise accuracy correlations with humans are weak and non-significant for all models except ChatGPT 4o. This indicates that models might have learned abstract SI relationships to represent human biases of ranking types beyond memorizing item-wise frequency patterns.

6.2 Extra Semantic Information

In this experiment, we replaced the **neighbor** information in the previous prompt with additional information about the antonymic **semantic** information. This took the form of '*the word* [A] *and the word* [B] *are antonyms*' (e.g., '*A = "cheap" and B = "luxurious" are antonyms*'). Similarly to the Extra Task Information experiment, we ran three randomly initialized runs for each of three prompt templates to calculate the average performances reported below.

6.2.1 Comparable Human Performance

The overall accuracy of 76% and error patterns by humans were comparable to baseline human performance in the Just-Ranking task, suggesting that emphasizing the antonymic relationships between adjectives does not significantly affect human rankings. We found no significant correlation between item frequency and accuracy (Spearman $\rho = 0.07$), though we observe that the extra semantic information boosts the accuracy for the less frequent items more than for the more frequent items with a significance that was merely a trend ($p = 0.07$), likely due to the overall low accuracy gain (see Figure 8 in Appendix C).

6.2.2 LLM Performance Drop

Eight out of ten models exhibited degraded accuracy from -8.66% to -41.84%, including Gemini 2.5 Pro which outperformed humans in the previous two tasks. Even after the drop, Gemini 2.5 Pro still achieved a higher accuracy than humans. The two models which improved in this condition (Qwen3-245B and ChatGPT 4o) were still far below human performance (41.2% and 18.24%).

Similarly to the experiment with extra task information, all ten models expanded their coverage of invalid rankings compared to the baseline Just Ranking experiment, illustrating that the extra information functioned as noise to shift model representations towards considering more types of wrong rankings. We find significant strong positive correlations between models and human for both all-answer and wrong-answer distribution. This shows that despite the distraction, models tend to group words into concatenated single SI pairs, similar to humans.

In contrast to the extra task information experiment, five models exhibited moderate positive item-wise correlation with human accuracy when both humans and models are provided with extra antonymic semantic information (Qwen3-235B, Qwen3-30B, ChatGPT o4-mini-high, Claude 3.7, and Deepseek-Prover-v2). This difference hints that while both kinds of extra information lowered models' performances, the distraction mechanisms diverge for these five models. Namely, extra semantic information causes models to better align item-wise to humans while extra task information does not.

6.2.3 Discussion

We have shown that all models understanding of cross-scale adjectives can be disrupted by extra task or semantic information that does not impede human performances on the same tasks. In both conditions with extra information, 60-80% of tested models exhibited degraded ranking accuracy but their correlation to human ranking patterns greatly increased. When models produce more invalid rankings, they tend to align with the humans bias to simply concatenate the two individual SIs.

While the extra task information reduces the item-wise correlation between humans and most tested models, the extra semantic information improves item-wise correlation with humans for five models. This contrast suggests that models did treat two kinds of additional information as different noises.

We also find that while human accuracy on less frequent items experienced a higher boost with extra information than more frequent items, many model accuracies of more frequent items dropped more than those of less frequent items (See Figure 12 in Appendix C). Such deviation underscores that

these models failed to encode the SIs or use the extra information in a human-like manner and that the disruption effects caused by the extra information were not countered by item frequencies.

6.3 Within-model consistency across ranking conditions

To examine item-level effects from extra information, we compared each model's ranking for each item in the baseline condition to its ranking in the Extra Task / Semantic information condition in terms of validity. We labeled the model's ranking for each item under each condition as valid if the correct ranking occurs in majority among multiple trials under the same experiment condition. We used four types of transitions (*Valid/Invalid→Valid/Invalid*) to categorize the model's performance on tested items. Ten items (∼10% of total items) in the *Valid→Invalid* category for *ChatGPT 4o* under the *Just-Ranking → Extra Task Info* comparison means that *ChatGPT 4o* ranked these items correctly in the baseline but not when extra Task information was provided (see Figure 5).

By comparing items in each transition type of each mode across two extra information conditions, we found that half of the lower-performance models in Just Ranking (*Qwen3-235B, Qwen3-30B, ChatGPT 4o, Gemini 2.0 Flush, Deepseek-Prover-V2*) have high overlap percentages of items across conditions for both *Valid→Invalid* and *Invalid→Invalid* categories, this reinforces that contrary to humans, extra information does not ease the difficulty of ranking for models. The other better-performing half (*Qwen2.5 Max, ChatGPT o4-mini-high, Gemini 2.5 Pro, Deepseek-R1, Claude 3.7*) has high overlap across conditions for both *Invalid→Valid* and *Valid→Valid* (Figure 13 in Appendix H). This pattern demonstrates that on an item-level, models' accuracy scores do not exhibit strong sensitivity to the difference between task and semantic information. In addition, models of the same family have near zero item overlap across transition types, underscoring that model family difference could not sufficiently explain item-level variations among models (see Appendix I).

We also noticed models vary their wrong rankings' choice range by different items. For items that models failed in both Just Ranking and Extra Info conditions (the *Invalid→Invalid* transition), models tended to give different wrong rankings instead of a dominant wrong one (Table 14 in Appendix J). Items in the *Invalid→Valid* transition types under both extra information condition had a clear majority wrong ranking in the baseline, implying for those items the models had strong but incorrect prior beliefs that the extra information helped correct (Figure 17 in Appendix J). For *Valid→Invalid* transitions, most models are more likely to produce a dominant wrong ranking for an item in the Extra Semantic Info condition than the Extra Task Info (Figure 15 in Appendix J), mirroring previous findings that models treat two kinds of extra info as different noises.

6.4 Generation & Judgment (In)Consistency of LLMs

To evaluate the consistency between generation and comprehension, we further compared eight of the tested models.[4] We generated invalid rankings for each item according to common invalid ranking types. In this task, we asked the models to judge whether they agreed with the validity of a given ranking. For each judgment prompt (see example prompt in Table 6 at Appendix B), we mixed valid/invalid rankings of items with varying proportions and asked the models to judge whether they agree with the given rankings. Only one ranking per item is given in a single prompt. We performed 27 trials for each model to calculate the average judgment accuracy. We calculated judgment accuracies by awarding a point whenever a model identifies a valid ranking as correct or recognizes an invalid ranking as incorrect.

We found that not only do all eight models have low accuracies for correct rankings, they also perform worse at the four dominant error patterns of humans than other invalid ranking types (see Figure 11 in Appendix E). The low judgment accuracies for correct rankings echo the finding that model accuracy drops with extra task and semantic information, underscoring the inadequate robustness of cross-scale adjectival understanding. That said, the high correlation of model error patterns with humans also reflects a bias to treat the dominant invalid rankings preferred by humans as more acceptable than other invalid rankings.

7 General Discussion

Among seven models surpassing human performances on the Just-Ranking experiment for cross-

[4]The two missing models were discontinued during the time of this experiment.

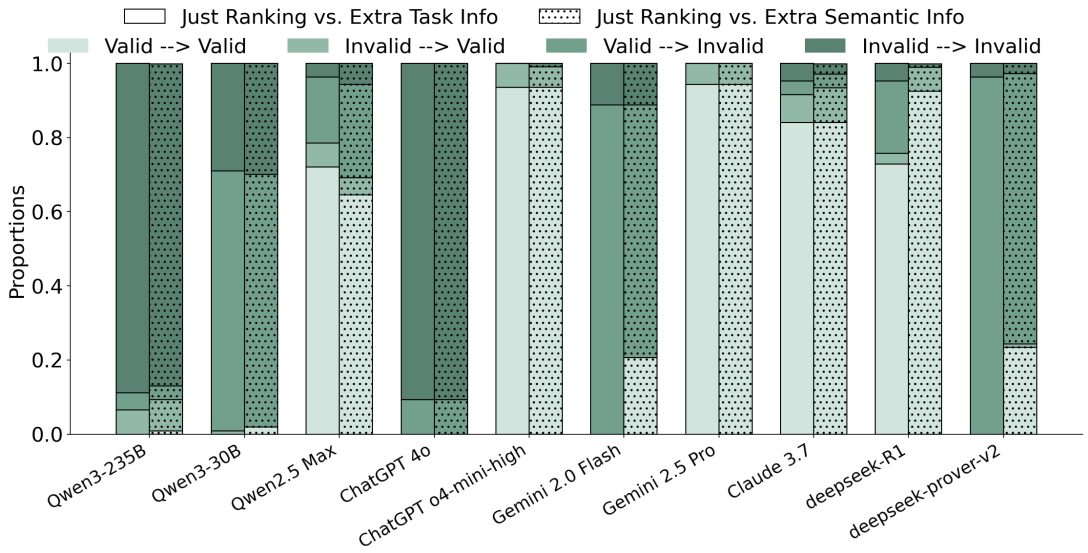

Figure 5: Item-wise ranking validity comparison between the Just Ranking baseline and Extra Task Information (left) / Extra Semantic Information (right). Arrows denote performance changes from the Just Ranking baseline to the extra-information conditions (e.g., Valid (Just Ranking) → Valid (Extra Info)).

scale adjectives, only Gemini 2.5 Pro maintained super-human performances on the same ranking tasks with additional information. And even Gemini 2.5 Pro lowered its performance when provided with extra semantic information that is simply a restatement of a lexically encoded antonym relationship between two words.

By comparing models within the same family, we found that bigger model size does not lead to better performance on this task (i.e. Qwen3-235B performed worse than Qwen3-30B). Since all tested models were trained on large amounts of data, they are expected to have high familiarity with items in the CrosSing dataset. Indeed, most models have strong correlations with humans in terms of kinds of produced rankings, biased towards invalid rankings where SI scales are simply concatenated. Such bias is further evidenced in our generation experiment where models commonly fail to recognize dominant invalid human ranking patterns as unacceptable.

Despite exhibiting human-like distributions of ranking patterns, half of the tested models had no significant item-wise correlation with humans in the Just-Ranking experiment and Ranking with Extra Semantic Information experiment. Nine out of ten models had no human-like item-sensitivity for the Ranking with Extra Task Information experiment. This gap in item-sensitivity indicates that these models use their representation of the gradable adjectives differently from humans in the ranking tasks. This finding was further echoed by model insensitivity to item frequency in the ranking experiments; while humans utilized extra information to boost accuracy of low frequency items more than high frequency ones, model performance degraded for high frequency items more than for low frequency items.

8 Conclusion

This study demonstrates that although most models we tested can rank cross-scale adjectives better than humans in a simple setting, their understanding of adjectival relationships is brittle and misaligned with humans. This fragility of model representation is evidenced by big accuracy drops in the presence of additional information that does not interfere with or which actively aids human reasoning. While these models produce human-like distributions of ranking patterns, they do not correlate with item-level human responses.

As the gradable adjectives studied in this paper have median frequency at the 89th percentile of the English vocabulary, if model understanding on these fundamental orderings does not align with humans, then problems dependent on the accurate usage of these adjectives will naturally emerge. This work provides a first step in better understanding model representations of cross-scale reasoning in hopes that future work can investigate methods of better aligning models with human responses.

Limitations

This work only focuses on English gradable adjectives, so the generalizability of models' fragility with extra information remains to be explored. While the gradable adjectives tested in **CrosSing** are of high frequency, the exact training frequency of words remains unknown due to the closed-source nature of the tested models' training data.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Anthropic. 2025. Claude 3.7 Sonnet System Card. Technical report, Anthropic. Accessed: 2025-10-06.

Helena Aparicio and Eszter Ronai. 2024. Scalar implicature rates vary within and across adjectival scales. *Proceedings of Semantics and Linguistic Theory 33*.

Helena Aparicio and Eszter Ronai. 2025. Scalar implicature rates vary within and across adjectival scales. *Journal of Semantics*, page ffaf002.

Andrea Beltrama and Ming Xiang. 2013. Is 'good' better than 'excellent'? an experimental investigation on scalar implicatures and gradable adjectives. In *Proceedings of sinn und bedeutung*, volume 17, pages 81–98.

Lewis Bott and Ira A. Noveck. 2004. Some utterances are underinformative: The onset and time course of scalar inferences. *Journal of memory and language*, 51(3):437–457.

Marc Brysbaert, Boris New, and Emmanuel Keuleers. 2012. Adding part-of-speech information to the subtlex-us word frequencies. *Behavior research methods*, 44(4):991–997.

Ye-eun Cho and Seong mook Kim. 2024. Pragmatic inference of scalar implicature by llms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 10–20.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Judith Degen. 2015. Investigating the distribution of some (but not all) implicatures using corpora and web-based methods. *Semantics and Pragmatics*, 8:11–1.

Judith Degen and Michael K. Tanenhaus. 2015. Processing scalar implicature: A constraint-based approach. *Cognitive science*, 39(4):667–710.

Judith Degen and Michael K. Tanenhaus. 2016. Availability of alternatives and the processing of scalar implicatures: A visual world eye-tracking study. *Cognitive science*, 40(1):172–201.

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.

Michael C. Frank and Noah D. Goodman. 2012. Predicting pragmatic reasoning in language games. *Science*, 336(6084):998.

Noah D. Goodman and Michael C. Frank. 2016. Pragmatic language interpretation as probabilistic inference. *Topics in cognitive science*, 20(11):818–829.

Noah D. Goodman and Andreas Stuhlmüller. 2013. Knowledge and implicature: Modeling language understanding as social cognition. *Topics in cognitive science*, 5(1):173–184.

Nicole Gotzner, Stephanie Solt, and Anton Benz. 2018. Scalar diversity, negative strengthening, and adjectival semantics. *Frontiers in psychology*, 9:1659.

Herbert P. Grice. 1975. Logic and conversation. In *Speech acts*, pages 41–58. Brill.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Laurence R. Horn. 1989. *A natural history of negation*. University of Chicago Press.

Jennifer Hu, Roger Levy, Judith Degen, and Sebastian Schuster. 2023. Expectations over unspoken alternatives predict pragmatic inferences. *Transactions of the Association for Computational Linguistics (TACL)*, 11:885—901.

Jennifer Hu, Roger Levy, and Sebastian Schuster. 2022. Predicting scalar diversity with context-driven uncertainty over alternatives. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, page 68–74.

Paloma Jeretic, Alex Warstadt, Suvrat Bhooshan, and Adina Williams. 2020. Are natural language inference models imppressive? learning implicature and presupposition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8690–8705.

Alyssa F. Kampa, Anna Papafragou, and Kaja K. Jasińska. 2025. Scalar inference is supported by theory of mind networks in adults and children. *Language, Cognition and Neuroscience*, pages 1–21.

Napoleon Katsos and Dorothy V.M. Bishop. 2011. Pragmatic tolerance: Implications for the acquisition of informativeness and implicature. *Cognition*, 120(1):67–81.

Joo-Kyung Kim and Marie-Catherine de Marneffe. 2013. Deriving adjectival scales from continuous space word representations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1625–1630.

Stephen C. Levinson. 2000. *Presumptive meanings: The theory of generalized conversational implicature*. MIT press.

Rashid Nizamani, Sebastian Schuster, and Vera Demberg. 2024. Siga: A naturalistic nli dataset of english scalar implicatures with gradable adjectives. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14784–14795.

Ira A. Noveck. 2001. When children are more logical than adults: Experimental investigations of scalar implicature. *Cognition*, 78(2):165–188.

OpenAI. 2025. OpenAI o3 and o4-mini System Card. Technical report, OpenAI. Accessed: 2025-10-06.

Anna Papafragou and Julien Musolino. 2003. Scalar implicatures: experiments at the semantics–pragmatics interface. *Cognition*, 86(3):253–282.

Alicia Parrish, Sebastian Schuster, Alex Warstadt, Omar Agha, Soo-Hwan Lee, Zhuoye Zhao, Samuel R. Bowman, and Tal Linzen. 2021. Nope: A corpus of naturally-occurring presuppositions in english. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 349–366.

Nausicaa Pouscoulous, Ira A. Noveck, Guy Politzer, and Anne Bastide. 2007. A developmental investigation of processing costs in implicature production. *Language acquisition*, 14(4):347–375.

Zhuang Qiu, Xufeng Duan, and Zhenguang Garry Cai. 2023. Pragmatic implicature processing in chatgpt. *PsyArXiv Preprints*.

ZZ Ren, Zhihong Shao, Junxiao Song, Huajian Xin, Haocheng Wang, Wanjia Zhao, Liyue Zhang, Zhe Fu, Qihao Zhu, Dejian Yang, and 1 others. 2025. Deepseek-prover-v2: Advancing formal mathematical reasoning via reinforcement learning for subgoal decomposition. *arXiv preprint arXiv:2504.21801*.

Eszter Ronai and Ming Xiang. 2022. Three factors in explaining scalar diversity. In *Proceedings of sinn und bedeutung*, volume 22, pages 716–733.

Eszter Ronai and Ming Xiang. 2023. Tracking the activation of scalar alternatives with semantic priming. In *Experiments in Linguistic Meaning*, volume 2, pages 229–240.

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 System*, volume 36, pages 20827–20905.

Sebastian Schuster, Yuxing Chen, and Judith Degen. 2020. Harnessing the linguistic signal to predict scalar inferences. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5387–5403.

Dimitrios Skordos and Anna Papafragou. 2016. Children's derivation of scalar implicatures: Alternatives and relevance. *Cognition*, 153:6–18.

Aina Garí Soler and Marianna Apidianaki. 2020. Bert knows punta cana is not just beautiful, it's gorgeous: Ranking scalar adjectives with contextualised representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7371—-7385.

Settaluri Sravanthi, Meet Doshi, Pavan Tankala, Rudra Murthy, Raj Dabre, and Pushpak Bhattacharyya. 2024. Pub: A pragmatics understanding benchmark for assessing llms' pragmatics capabilities. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12075–12097.

John R. Starr, Ashlyn Winship, and Marten van Schijndel. 2025. Generating representations in space with GRIS. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 47, pages 1584–1590.

Bob van Tiel, Emiel Van Miltenburg, Natalia Zevakhina, and Bart Geurts. 2016. Scalar diversity. *Journal of semantics*, 33(1):137–175.

Bob van Tiel and Elizabeth Pankratz. 2021. Adjectival polarity and the processing of scalar inferences. *Glossa: a journal of general linguistics*, 6(1):32.

Bob van Tiel, Elizabeth Pankratz, and Chao Sun. 2019. Scales and scalarity: Processing scalar inferences. *Journal of Memory and Language*, 105:93–107.

Bob van Tiel and Walter Schaeken. 2017. Processing conversational implicatures: alternatives and counterfactual reasoning. *Cognitive science*, 41:1119–1154.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

A Model License and Parameters

Models	License	Parameters	Context
ChatGPT 4o	OpenAI's Terms of Use for Free Users	Unknown	128k tokens
ChatGPT o4-mini-high	OpenAI's Terms of Use for Free Users	Unknown	200k tokens
Gemini 2.5 Pro	Gemini API Terms of Service	Unknown	1M tokens
Gemini 2.0 Flash	Gemini API Terms of Service	Unknown	1M tokens
Claude 3.7	Anthropic's Commercial Terms of Service	Unknown	200k tokens
Qwen2.5 Max	Alibaba Commercial Terms of Usage	Unknown	32k tokens
Qwen3-235B	Apache 2.0	235B	131k tokens
Qwen3-30B	Apache 2.0	31.1B	131k tokens
deepseekR1	MIT	685B	128k tokens
deepseek-prover-v2	deepseek-license	685B	unknown

Table 4: Model licenses and parameters.

B How LLMs generate and judge scales

Pure-Generation1.7k tokens	Rank the words on a scale within each list. [good, bad, excellent, terrible]...(109 items) ...(109 items)
NC-SG 1.7k tokens	The first word and the second word should be positioned next to each other in the ranking. Rank the words on a scale within each list. [good, bad, excellent, terrible] ...(109 items in default order)
NC-RP-POS 3.5k tokens	Rank the words on a scale within each list. [good, bad, excellent, terrible] should be ranked with first word and the third word next to each other. ...(109 items)
NC-RP-Word 3.7k tokens	Rank the words on a scale within each list. [good, bad, excellent, terrible] should be ranked with 'good' and 'excellent' next to each other. ...(109 items)
AC-SG 1.7k tokens	The first word and the third word are antonyms. Rank the words on a scale within each list. [good, bad, excellent, terrible] ...(109 items in default order)
AC-RP-POS 3.5k tokens	Rank the words on a scale within each list. [good, bad, excellent, terrible] should be ranked given that the third word and the fourth word are antonyms. ...(109 items)
AC-RP-Word 3.7k tokens	Rank the words on a scale within each list. [good, bad, excellent, terrible] should be ranked given that 'excellent' and 'terrible' are antonyms. ...(109 items)

Table 5: Prompts for generation tasks. Three variables were manipulated to examine how SOTA LLMs handle constrained scalar reasoning: (1) constraint type (NC/AC), (2) constraint frequency (SG/RP), and (3) referring expression type for the target words (POS/Word). SG = singular constraint occurrence; RP = repeated constraint occurrence; POS = position-based word references. Token numbers indicate average one-shot prompt length.

Judgment	Are the following lists valid orderings to reflect the appropriate relations in each scale? [good, bad, excellent, terrible]...(109 items) ...(109 items) [0.5k tokens]

Table 6: Prompts for judgment tasks.

Models	Pure Generation	Generation with NCs	Generation with ACs
ChatGPT 4o	0.076453	0.179409	0.182467
ChatGPT o4-mini-high	0.944954	0.876656	0.740061
Gemini 2.5 Pro	0.948012	0.958206	0.806575
Gemini 2.0 Flash	0.828746	0.188583	0.497961
Claude 3.7	0.788991	0.636086	0.702345
Qwen2.5 Max	0.694190	0.593612	0.546891
Qwen3-235B	0.155963	0.354400	0.412063
Qwen3-30B	0.657492	0.288141	0.321865
deepseekR1	0.853211	0.545702	0.750255
deepseek-prover-v2	0.886850	0.282025	0.468400

Table 7: Overall model performance accuracy scores in all generation tasks. NC refers to Extra Task Information, and AC refers to Extra Semantic Information.

ChatGPT 4o	N/A	ChatGPT o4-mini-high	**0.913**
Gemini 2.5 Pro	0.7854	Gemini 2.0 Flash	N/A
Claude 3.7	0.7735	Qwen2.5 Max	0.8233
Qwen3-235B	**0.6504**	Qwen3-30B	0.7808
deepseekR1	0.7621	deepseek-prover-v2	N/A

Table 8: Average Accuracy of Ranking Judgment Task (Maximum Accuracy = 1)

C Frequency Correlations Across Tasks

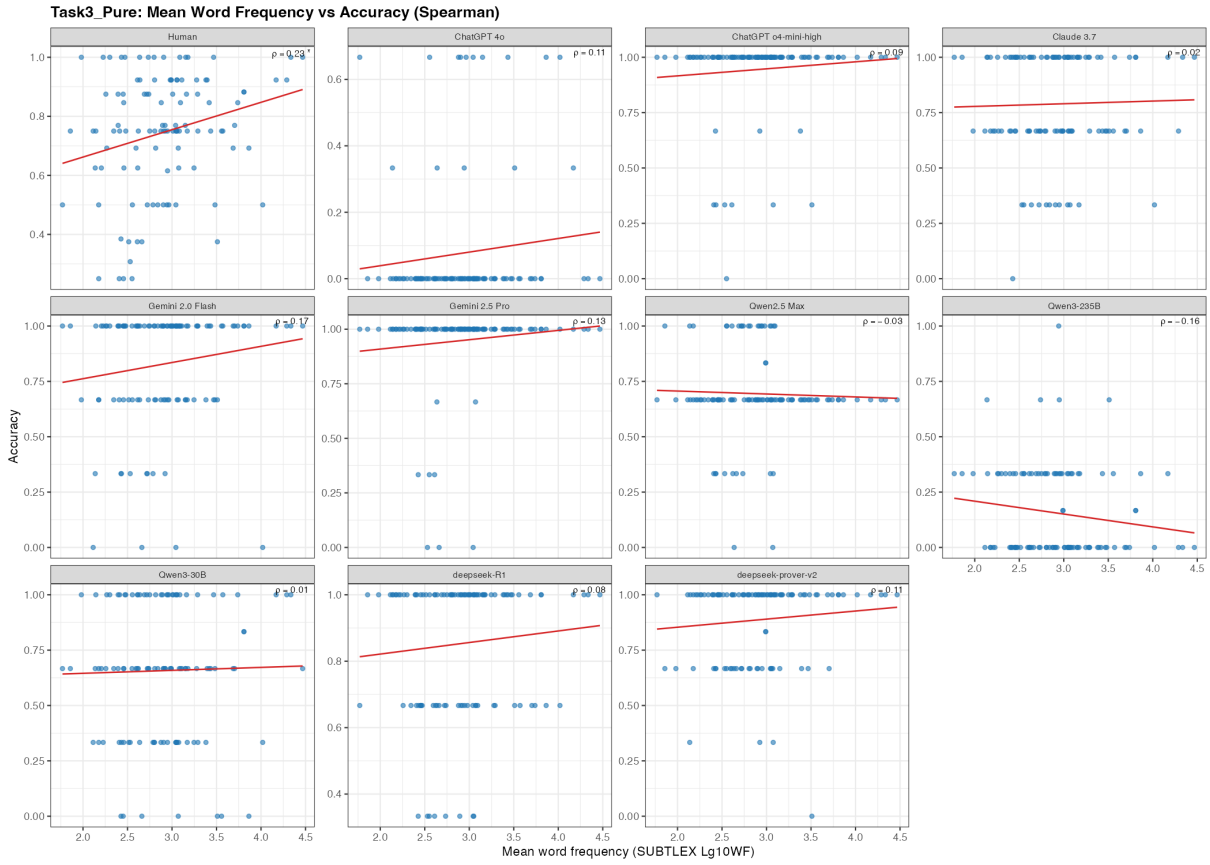

Figure 6: Just-Ranking Experiment: Correlation between Average Word Frequency and Ranking Accuracy of Items

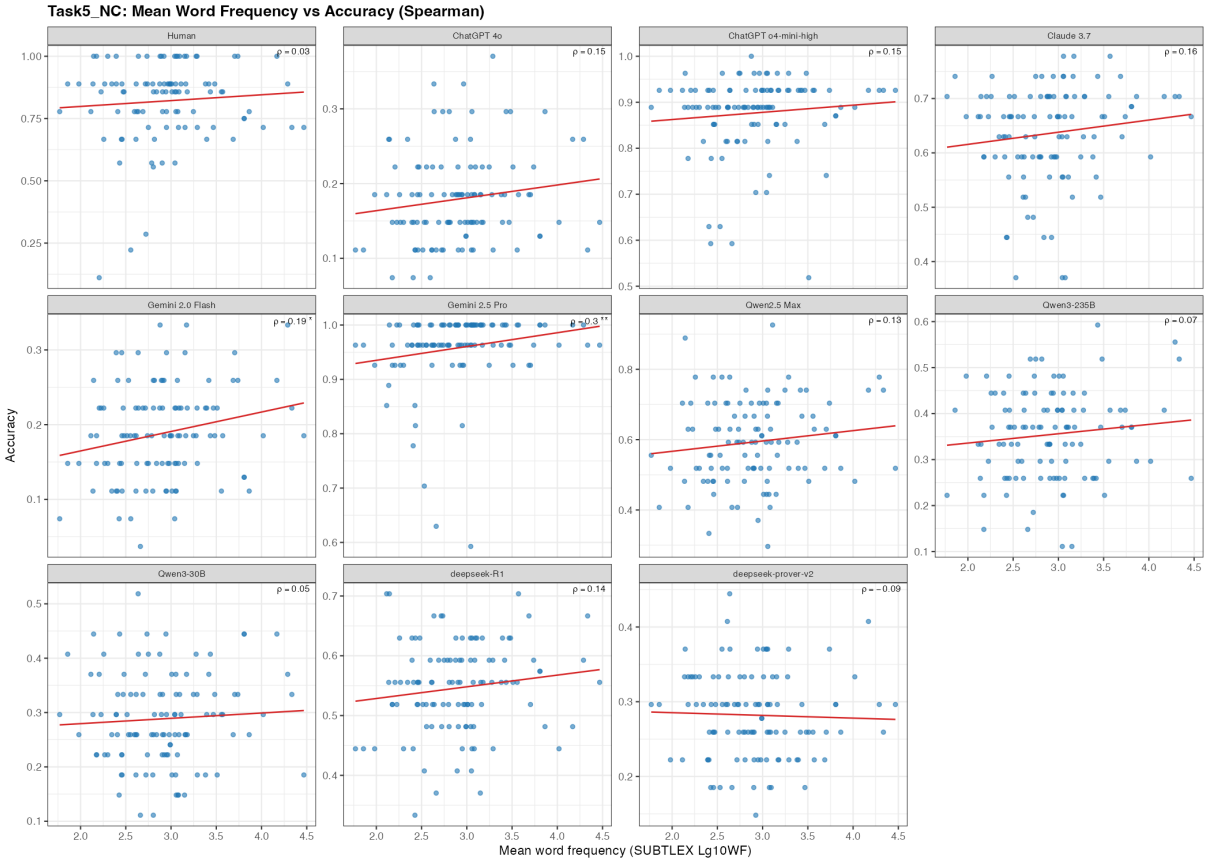

Figure 7: Ranking with Extra Task Information Experiment: Correlation between Average Word Frequency and Ranking Accuracy of Items. NC refers to Extra Task Information.

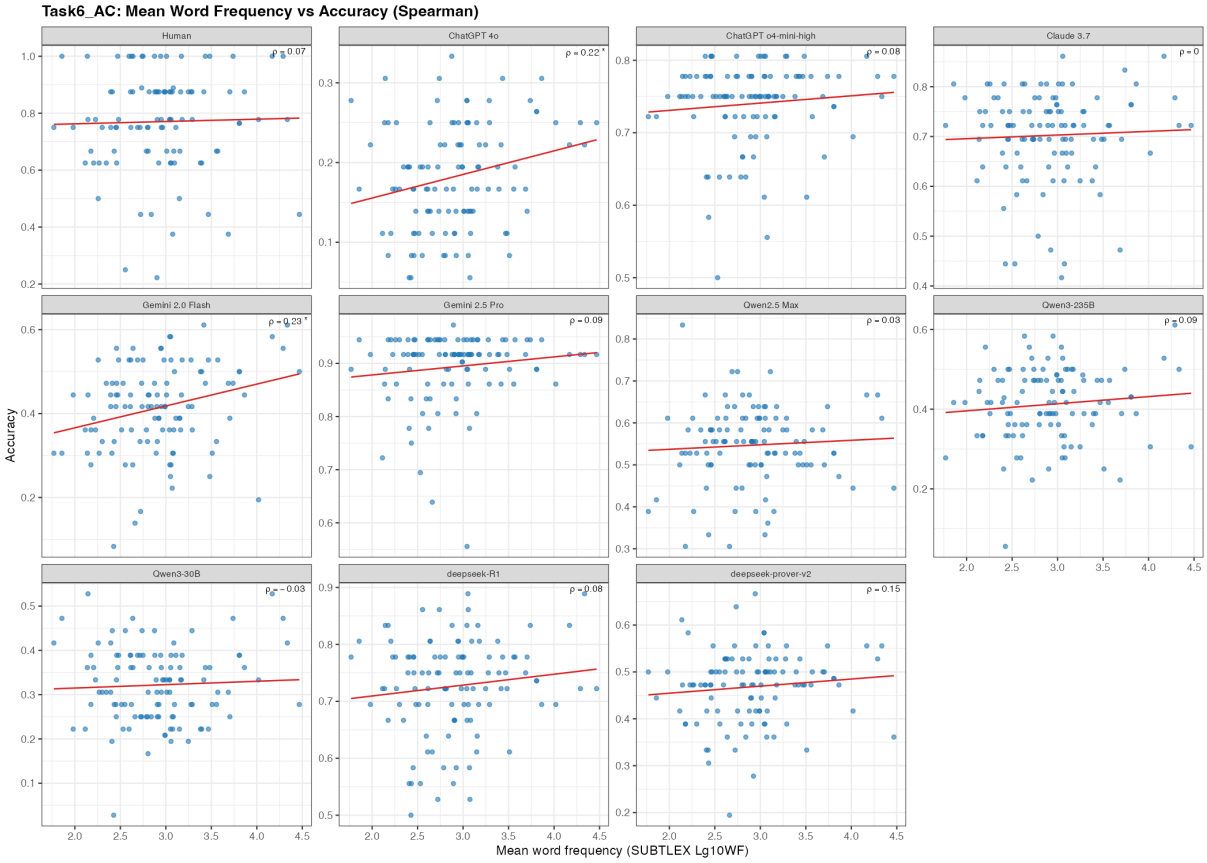

Figure 8: Ranking with Extra Semantic Information Experiment: Correlation between Average Word Frequency and Ranking Accuracy of Items. AC refers to Extra Semantic Information experiment.

D Correlation Analysis

Model	SC	SC_N	SC_A
Qwen3-235B	0.29	0.74^{***}	0.72^{***}
Qwen3-30B	0.5^{*}	0.67^{***}	0.44^{*}
Qwen2.5 Max	0.37	0.73^{***}	0.52^{*}
ChatGPT 4o	0.61^{**}	0.74^{***}	0.77^{***}
o4-mini-high	0.71^{**}	0.76^{***}	0.78^{***}
Gemini 2.0 Flash	0.41	0.64^{**}	0.72^{***}
Gemini 2.5 Pro	0.72^{***}	0.34	0.71^{**}
Claude 3.7	0.59^{**}	0.76^{***}	0.68^{**}
Deepseek-R1	0.57^{*}	0.57^{*}	0.59^{**}
DS-Prover-v2	0.7^{***}	0.7^{***}	0.59^{**}

Table 9: Correlation with Human Invalid Patterns for All Experiments: SC denotes each model's response pattern's spearman correlation to the human baseline. The * symbol denotes p-value thresholds where '***' is $p < 0.001$, '**'< 0.01, and '*'< 0.05. '$_N$' means the metric values are for Extra Task Information experiment, and '$_A$' refers to the Extra Semantic Information experiment. 'o4-mini-high' stands for ChatGPT o4-mini-high and 'DS-Prover-v2' for 'Deepseek-Prover-v2'.

E Barplots for Models Generation Answer types

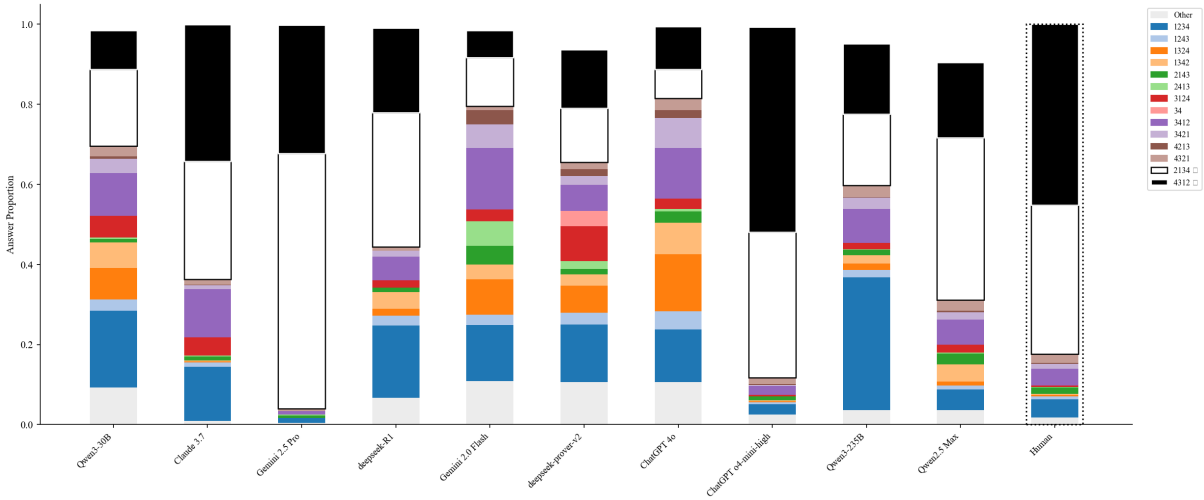

Figure 9: Ranking with Extra Task Info - Response Patterns of LLMs and Humans. Each color bar corresponds to a ranking type.

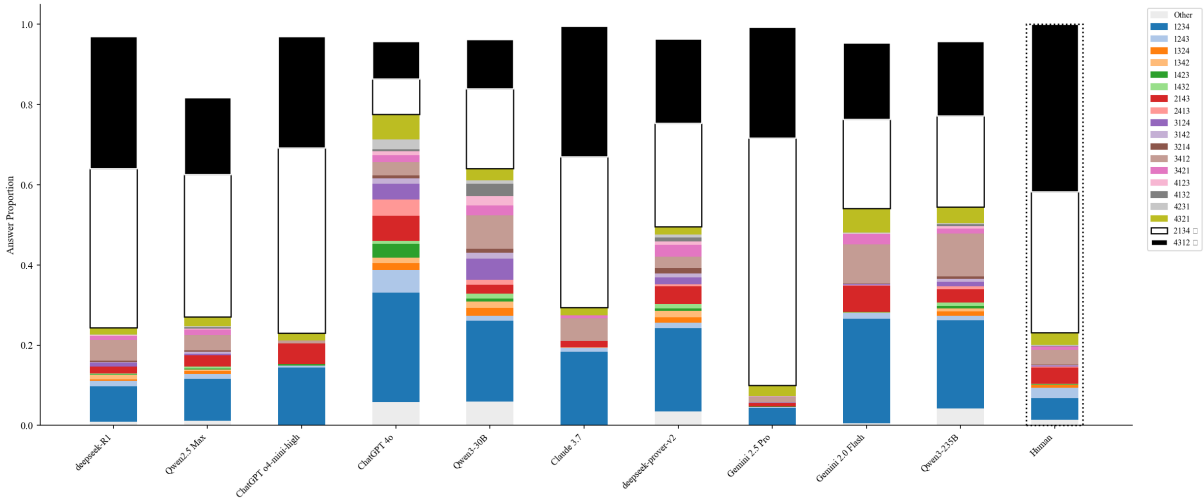

Figure 10: Ranking with Extra Semantic Info - Response Patterns of LLMs and Humans. Each color bar corresponds to a ranking type.

F LLMs Judgment Accuracy

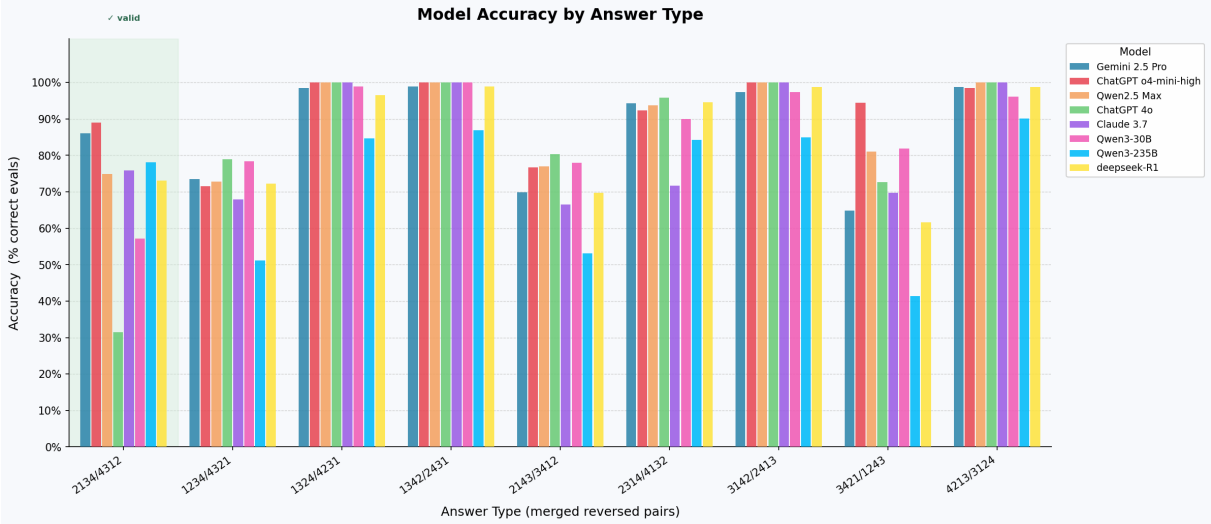

Figure 11: Ten LLMs Judgment Accuracies by Answer Types

G Frequency x Extra Information Interaction

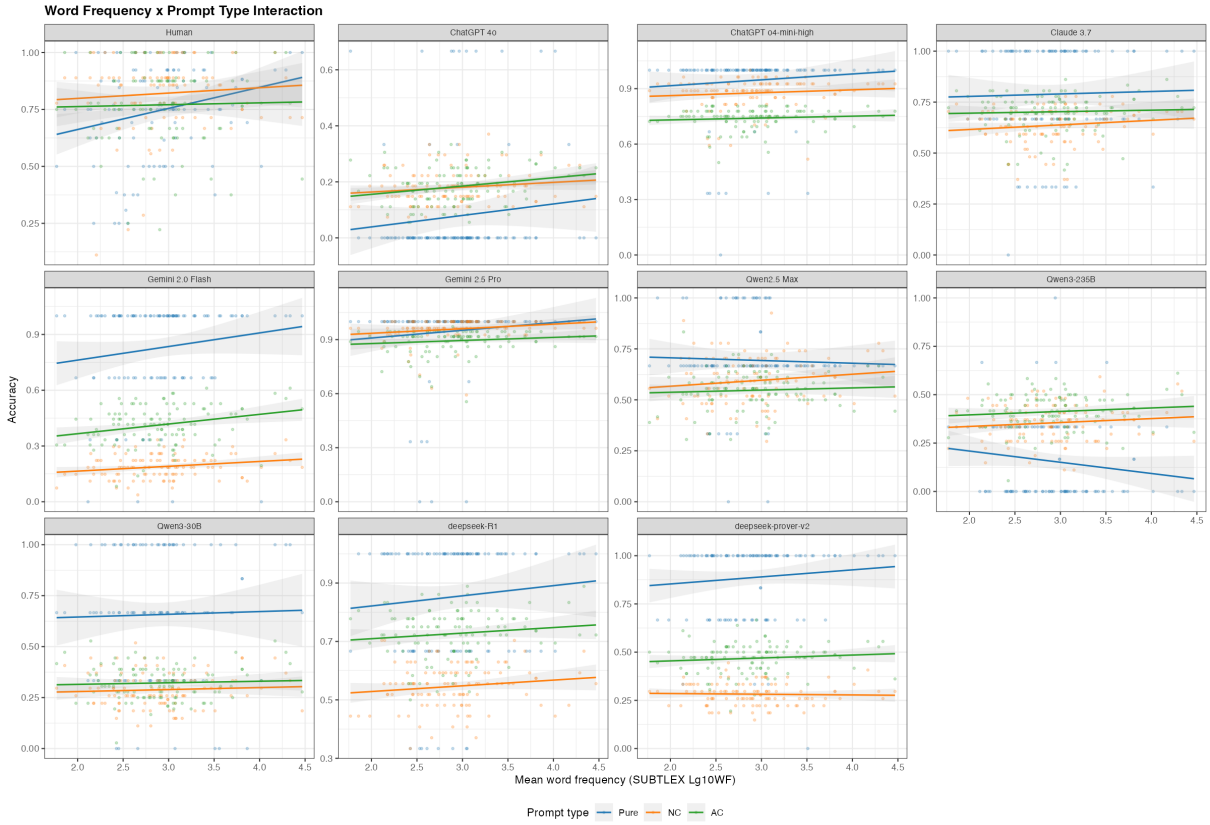

Figure 12: Ten LLMs and human interaction with frequency and extra information. Pure refers to Just-Ranking experiment; NC refers to Extra Task Information experiment; and AC refers to Extra Semantic Information experiment.

H Extra Task vs. Semantic Info

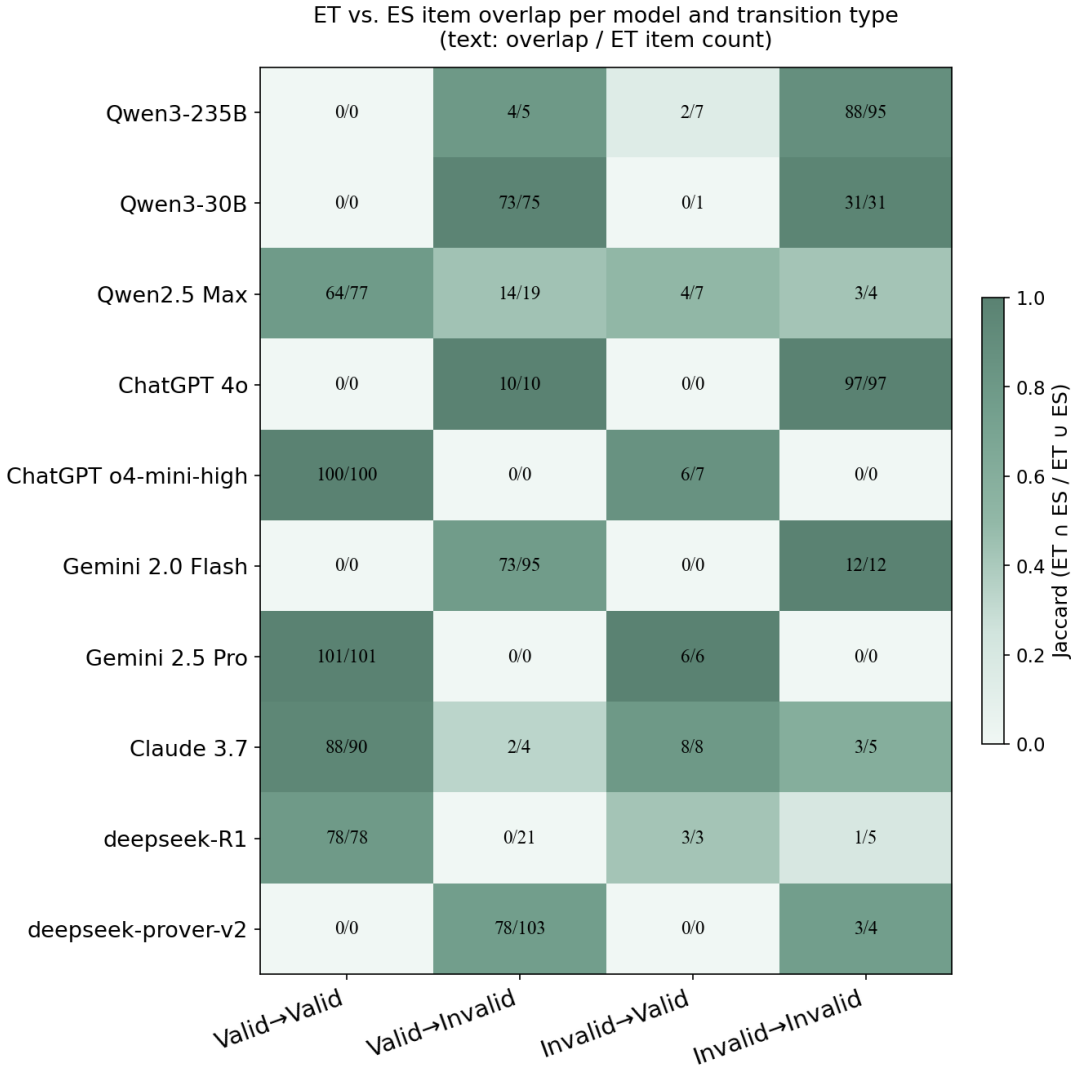

Figure 13: Heatmap of Jaccard Similarity per transition type and model between Extra Task (ET) and Extra Semantic (ES)'s comparison to the Just-Ranking baseline. Darker green means high similarity. The x/y ratio in each box means the number of items shared divided by the union of items in the transition category from both ET/ES transitions from the baseline.

I Within-Language-Model-Family Item Overlap

I.1 Extra Task Info (ET)

Family	Comparison	V→V	V→I	I→V	I→I
DeepSeek	R1 vs. prover-v2	0.00 (0/107)	0.19 (20/107)	0.00 (0/107)	0.00 (0/107)
Gemini	2.0 Flash vs. 2.5 Pro	0.00 (0/107)	0.00 (0/107)	0.00 (0/107)	0.00 (0/107)
ChatGPT	4o vs. o4-mini-high	0.00 (0/107)	0.00 (0/107)	0.00 (0/107)	0.00 (0/107)
Qwen	Q3-235B vs. Q3-30B	0.00 (0/107)	0.05 (4/107)	0.00 (0/107)	0.31 (30/107)
	Q3-235B vs. Q2.5 Mx	0.00 (0/107)	0.09 (2/107)	0.00 (0/107)	0.04 (4/107)
	Q3-30B vs. Q2.5 Mx	0.00 (0/107)	0.13 (11/107)	0.14 (1/107)	0.13 (4/107)

Table 10: Within-family item overlap for the **ET** condition (Just Ranking vs. Extra Task Info). Each cell shows Jaccard (overlap/107) where 107 is the total number of conceptual items per model. Rows within each family are all pairwise model comparisons. 'V' stands fro 'Valid' and 'I' stands for 'Invalid'.

I.2 Extra Semantic Info (ES)

Table 11: Within-family item overlap for the **ES** condition (Just Ranking vs. Extra Semantic Info). Each cell shows Jaccard (overlap/107) where 107 is the total number of list items per model. Rows within each family are all pairwise model comparisons.

Family	Comparison	V→V	V→I	I→V	I→I
DeepSeek	R1 vs. prover-v2	0.23 (23/107)	0.00 (0/107)	0.00 (0/107)	0.00 (0/107)
Gemini	2.0 Flash vs. 2.5 Pro	0.22 (22/107)	0.00 (0/107)	0.00 (0/107)	0.00 (0/107)
ChatGPT	4o vs. o4-mini-high	0.00 (0/107)	0.00 (0/107)	0.00 (0/107)	0.01 (1/107)
Qwen	Q3-235B vs. Q3-30B	0.00 (0/107)	0.04 (3/107)	0.00 (0/107)	0.30 (29/107)
	Q3-235B vs. Q2.5 Mx	0.01 (1/107)	0.07 (2/107)	0.08 (1/107)	0.07 (6/107)
	Q3-30B vs. Q2.5 Mx	0.03 (2/107)	0.19 (16/107)	0.00 (0/107)	0.15 (5/107)

Table 12: Within-family item overlap for the **ES** condition (Just Ranking vs. Extra Semantic Info). Each cell shows Jaccard (overlap/107) where 107 is the total number of list items per model. Rows within each family are all pairwise model comparisons. 'V' stands fro 'Valid' and 'I' stands for 'Invalid'.

J Baseline to Extra Information Item-level Validity Change Analyses

J.1 Valid→Valid

Model	Cond.	n	Both Maj.%	Changed%	Unchanged%	Ambig.%
Qwen3-235B	ET	30	80.0%	6.7%	73.3%	20.0%
	ES	32	53.1%	3.1%	50.0%	46.9%
Qwen3-30B	ET	17	52.9%	5.9%	47.1%	47.1%
	ES	60	31.7%	0.0%	31.7%	68.3%
Qwen2.5 Max	ET	63	76.2%	15.9%	60.3%	23.8%
	ES	105	47.6%	6.7%	41.0%	52.4%
ChatGPT 4o	ET	4	50.0%	0.0%	50.0%	50.0%
	ES	3	0.0%	0.0%	0.0%	100.0%
ChatGPT o4-mini-high	ET	285	76.8%	2.8%	74.0%	23.2%
	ES	208	74.5%	4.8%	69.7%	25.5%
Gemini 2.0 Flash	ET	6	50.0%	33.3%	16.7%	50.0%
	ES	161	64.0%	31.7%	32.3%	36.0%
Gemini 2.5 Pro	ET	295	73.9%	17.6%	56.3%	26.1%
	ES	297	78.5%	22.9%	55.6%	21.5%
Claude 3.7	ET	203	57.1%	2.5%	54.7%	42.9%
	ES	210	60.5%	3.3%	57.1%	39.5%
deepseek-R1	ET	81	32.1%	2.5%	29.6%	67.9%
	ES	219	79.0%	12.3%	66.7%	21.0%
deepseek-prover-v2	ET	33	0.0%	0.0%	0.0%	100.0%
	ES	157	44.6%	0.0%	44.6%	55.4%

Table 13: Valid→Valid analysis (toggle=DIFF: treat reversed rankings as **distinct**). For each model and comparison condition, n is the number of presentations classified as V→V; *Both Maj.%* is the percentage with a clear majority output ranking in *both* Just Ranking and perturbed runs; *Changed%* / *Unchanged%* are subsets of Both Maj.; *Ambig.%* is the remainder. ET = Just Ranking vs. Extra Task Info; ES = Just Ranking vs. Extra Semantic Info.

J.2 Invalid→Invalid

Model	ET: $J_{\text{same}}/J_{\text{diff}}$ $[n]$	ES: $J_{\text{same}}/J_{\text{diff}}$ $[n]$
Qwen3-235B	0.071/0.062 [175]	0.049/0.032 [156]
Qwen3-30B	0.148/0.111 [106]	0.159/0.115 [96]
Qwen2.5 Max	0.041/0.036 [64]	0.048/0.036 [68]
ChatGPT 4o	0.154/0.129 [287]	0.122/0.041 [287]
ChatGPT o4-mini-high	0.358/0.293 [8]	0.324/0.202 [9]
Gemini 2.0 Flash	0.121/0.074 [53]	0.183/0.108 [38]
Gemini 2.5 Pro	0.467/0.233 [5]	0.500/0.333 [2]
Claude 3.7	0.304/0.225 [20]	0.365/0.261 [21]
deepseek-R1	0.175/0.095 [39]	0.246/0.158 [18]
deepseek-prover-v2	0.141/0.086 [37]	0.192/0.159 [24]

Table 14: Invalid→Invalid ranking overlap between conditions. Each cell shows $J_{\text{same}}/J_{\text{diff}}$ $[n]$ where J_{same} is the mean Jaccard treating reversed rankings as identical, J_{diff} treats them as distinct, and n is the number of I→I presentations. Low Jaccard indicates the model gives different wrong rankings across conditions; high Jaccard indicates consistently wrong rankings. ET = Just Ranking vs. Extra Task Info; ES = Just Ranking vs. Extra Semantic Info.

J.3 Valid→Invalid

Model	ET (same)	ET (diff)	ES (same)	ES (diff)
Qwen3-235B	5/17 (29.4%)	2/17 (11.8%)	2/15 (13.3%)	0/15 (0.0%)
Qwen3-30B	11/188 (5.9%)	10/188 (5.3%)	17/145 (11.7%)	4/145 (2.8%)
Qwen2.5 Max	54/153 (35.3%)	40/153 (26.1%)	51/111 (45.9%)	13/111 (11.7%)
ChatGPT 4o	3/21 (14.3%)	3/21 (14.3%)	6/22 (27.3%)	3/22 (13.6%)
ChatGPT o4-mini-high	4/12 (33.3%)	2/12 (16.7%)	48/89 (53.9%)	22/89 (24.7%)
Gemini 2.0 Flash	4/254 (1.6%)	3/254 (1.2%)	42/99 (42.4%)	7/99 (7.1%)
Gemini 2.5 Pro	1/4 (25.0%)	0/4 (0.0%)	1/2 (50.0%)	0/2 (0.0%)
Claude 3.7	3/43 (7.0%)	1/43 (2.3%)	4/36 (11.1%)	2/36 (5.6%)
deepseek-R1	19/186 (10.2%)	12/186 (6.5%)	13/48 (27.1%)	3/48 (6.2%)
deepseek-prover-v2	7/245 (2.9%)	3/245 (1.2%)	25/121 (20.7%)	14/121 (11.6%)

Table 15: Valid→Invalid: percentage of presentations where a dominant (majority, $> 50\%$ of runs) invalid output ranking exists in the perturbed condition. Format: number of items judged as invalid with a dominant ranking / total number of items judged as invalid for each toggle setting (toggle=same when reversed ranking is treated as the same with its reversal counterpart; toggle=diff when reversed ranking like 4321 is treated as different from 1234). A high percentage implies the model is systematically misled to one specific wrong ranking. ET = Just Ranking vs. Extra Task Info; ES = Just Ranking vs. Extra Semantic Info.

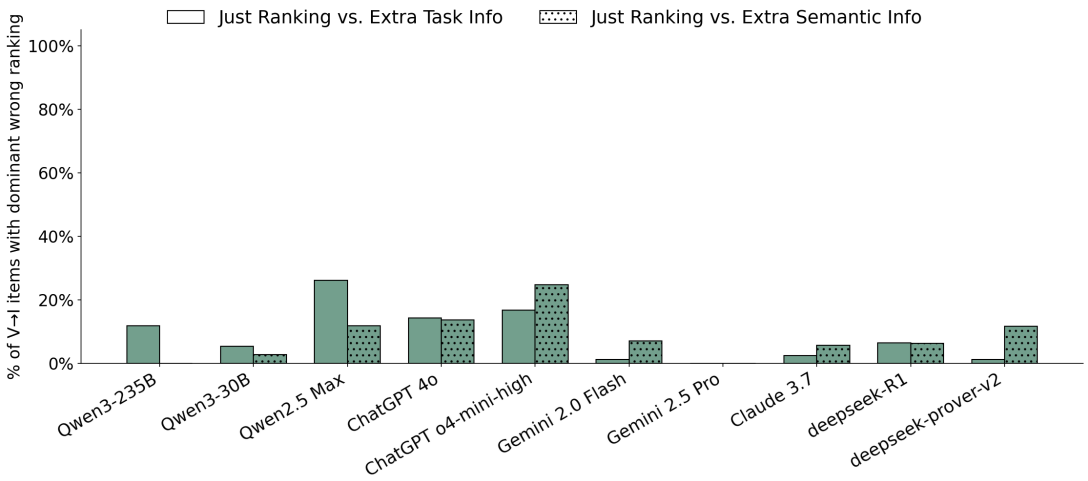

Figure 14: Extra Information Condition Comparison in terms of Valid→Invalid items per model, assuming reversal rankings are treated as different ($1234 \neq 4321$). The higher percentage means the model produced a specific dominant invalid rankings on more items in the extra information condition.

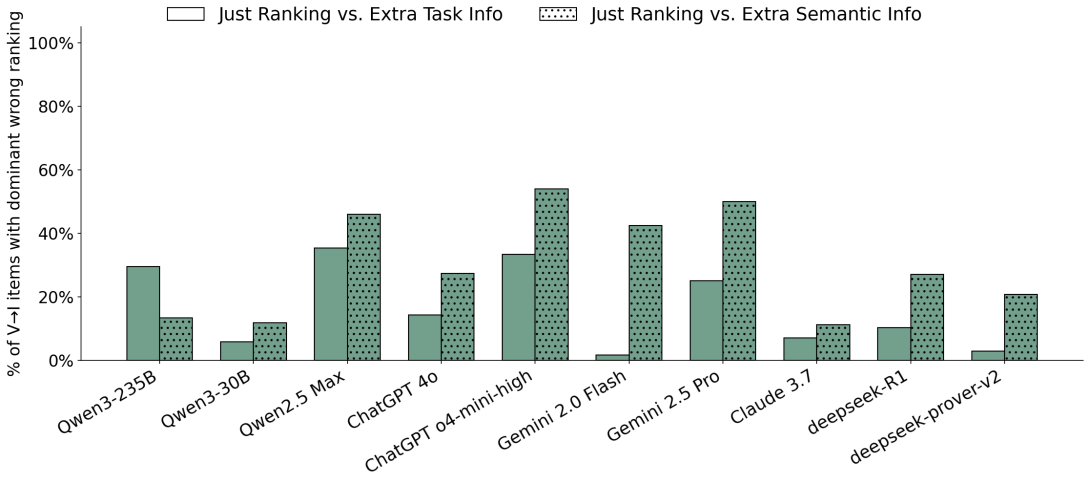

Figure 15: Extra Information Condition Comparison in terms of Valid→Invalid items per model, assuming reversal rankings are treated as the same ($1234 = 4321$). The higher percentage means the model produced a specific dominant invalid rankings on more items in the extra information condition.

J.4 Invalid→Valid

Model	ET (same)	ET (diff)	ES (same)	ES (diff)
Qwen3-235B	38/93 (40.9%)	38/93 (40.9%)	63/112 (56.2%)	63/112 (56.2%)
Qwen3-30B	3/4 (75.0%)	3/4 (75.0%)	12/14 (85.7%)	12/14 (85.7%)
Qwen2.5 Max	6/35 (17.1%)	6/35 (17.1%)	6/31 (19.4%)	6/31 (19.4%)
ChatGPT 4o	3/3 (100.0%)	3/3 (100.0%)	3/3 (100.0%)	3/3 (100.0%)
ChatGPT o4-mini-high	10/10 (100.0%)	10/10 (100.0%)	9/9 (100.0%)	9/9 (100.0%)
Gemini 2.0 Flash	0/2 (0.0%)	0/2 (0.0%)	12/17 (70.6%)	12/17 (70.6%)
Gemini 2.5 Pro	11/11 (100.0%)	11/11 (100.0%)	14/14 (100.0%)	14/14 (100.0%)
Claude 3.7	44/49 (89.8%)	44/49 (89.8%)	42/48 (87.5%)	42/48 (87.5%)
deepseek-R1	8/9 (88.9%)	8/9 (88.9%)	23/30 (76.7%)	23/30 (76.7%)
deepseek-prover-v2	–	–	10/13 (76.9%)	10/13 (76.9%)

Table 16: Invalid→Valid: percentage of presentations where the model had a dominant ($> 50\%$ of Just Ranking runs) invalid output ranking in the *Just Ranking* condition. Format: number of items judged as invalid with a dominant ranking in the baseline / total number of items judged as invalid in the baseline for each toggle setting (toggle=same when reversed ranking is treated as the same with its reversal counterpart; toggle=diff when reversed ranking like 4321 is treated as different from 1234). ET = Just Ranking vs. Extra Task Info; ES = Just Ranking vs. Extra Semantic Info.

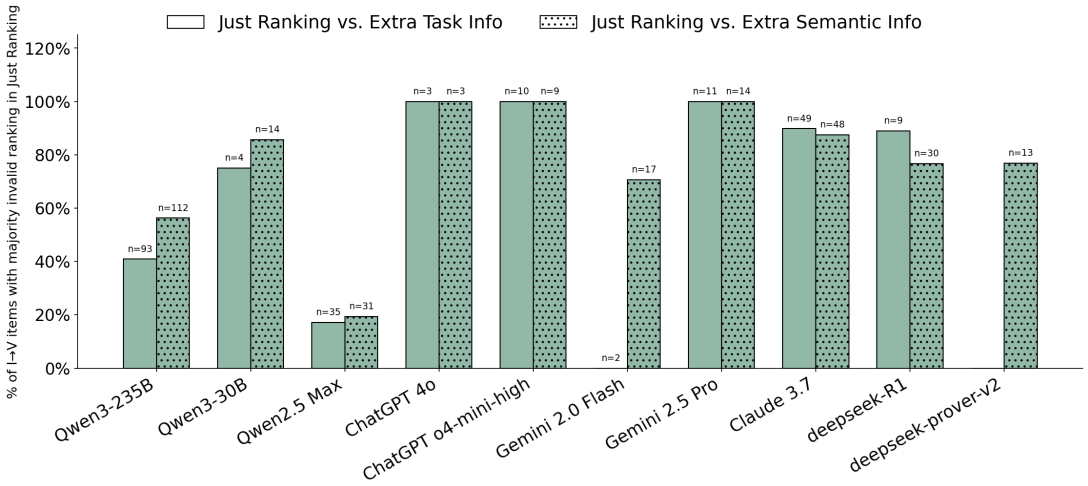

Figure 16: Extra Information Condition Comparison in terms of Invalid→Valid items per model, assuming reversal rankings are treated as different ($1234 \neq 4321$). The higher percentage means the model produced a specific dominant invalid rankings on more items in the baseline condition.

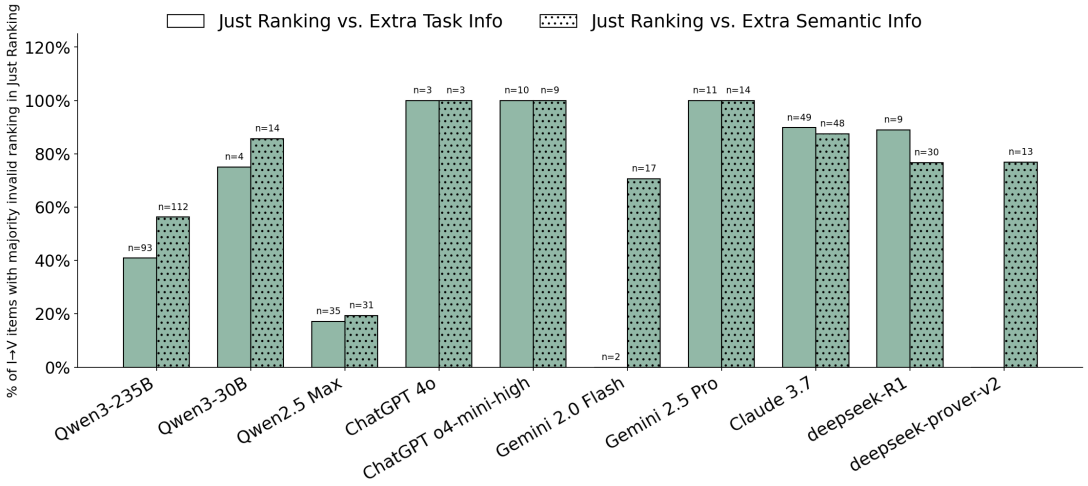

Figure 17: Extra Information Condition Comparison in terms of Invalid→Valid items per model, assuming reversal rankings are treated as the same ($1234 = 4321$). The higher percentage means the model produced a specific dominant invalid rankings on more items in the baseline condition.