Contrastive Decoding Mitigates Score Range Bias in LLM-as-a-Judge

Yoshinari Fujinuma
Patronus AI
fujinumay@gmail.com

Abstract

Large Language Models (LLMs) are commonly used as evaluators in various applications, but the reliability of the outcomes remains a challenge. One such challenge is using LLMs-as-judges for direct assessment, i.e., assigning scores from a specified range without any references. Using summarization as our primary testbed, we first show that this challenge stems from LLM judge outputs being associated with score range bias, i.e., LLM judge outputs are highly sensitive to pre-defined score ranges. We also show that similar biases exist among models from the same family. We then mitigate this bias through contrastive decoding, achieving up to 11.7% relative improvement in Spearman correlation with human judgments, averaged across score ranges.[1]

1 Introduction

Large Language Model (LLM) judges have become an integral component of the evaluation ecosystem (Lin et al., 2022; Chiang and Lee, 2023; Bubeck et al., 2023). In evaluations ranging from direct assessment, where judges evaluate individual outputs by assigning scores (Liu et al., 2023), to pairwise comparisons, where judges compare two outputs and determine which is superior (Zheng et al., 2023; Ye et al., 2024), using LLM as a judge is increasingly deployed to provide automatic, scalable, and cost-effective evaluation across diverse tasks. However, the reliability of such evaluations faces significant challenges, particularly when models assess their own outputs (Zheng et al., 2023) or those from the same model family (Goel et al., 2025). These biases constrain the set of models that can be reliably employed as LLM-as-a-judge for evaluation. But could there be any other biases hidden when using LLMs as judges?

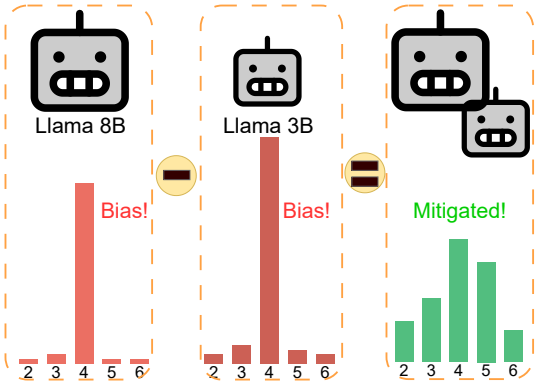


Figure 1: Overview of score range bias in 2-4 range and how contrastive decoding mitigates it through canceling out similar bias across models from the same family.

We reveal another bias in LLM judge outputs: **score range bias**, where LLM judges produce different correlations with human judgments under equivalent but shifted score ranges (e.g., 0-4 vs. 2-6), a phenomenon motivated by prior findings that LLMs struggle in simple arithmetic tasks (Nogueira et al., 2021; Gambardella et al., 2024). Upon identifying such biases, we also explore a mitigation strategy by connecting recent work on contrastive decoding (Li et al., 2023; O'Brien and Lewis, 2023) and family-enhancement bias (Goel et al., 2025), aiming to cancel out similar score range biases encoded across the models from the same family. We use summarization as our primary testbed, as prior work has shown LLM judges to be unreliable on summarization (Ye et al., 2024; Panickssery et al., 2024), and further validate that our findings generalize, a broad benchmark (Kim et al., 2025).

Our summarized contributions are as follows:

- We first show that LLM judges have *score range bias*, a bias observed across different model sizes and families (Llama-3 and Qwen-2.5) when judging on direct assessment.

[1]Code: https://github.com/akkikiki/contrastive_decoding_llm_judge

Findings of the Association for Computational Linguistics: ACL 2026, pages 13404–13418

- We then show that contrastive decoding, motivated by the observation of similar score range biases shared across models from the same family, successfully mitigates these biases.
- We show that these findings generalize beyond summarization to the diverse BigGen-Bench benchmark (§4.2).

2 Related Work

We now review the related work on LLM judges focusing on the judge tasks and their biases.

LLM Judge Tasks LLM judge tasks fall into two categories: *direct* assessment (Jones et al., 2024; Li et al., 2024; Zhu et al., 2025) and *pairwise* assessment (Zheng et al., 2023; Ye et al., 2024). Direct assessment (Liu et al., 2023) involves assigning numerical ratings to a single output example. In pairwise assessment, LLM judges show higher correlation with human preferences than direct assessment (Liu et al., 2024), supporting that the challenge remain in direct assessment, and we therefore focus on experimenting on direct assessment.

LLM Judge Biases One known bias in LLM judges is self-enhancement bias i.e., the tendency to favor their own output (Liu et al., 2023; Zheng et al., 2023; Ye et al., 2024) even in proprietary models like GPT-4 (Wataoka et al., 2024). Extending beyond self-enhancement bias, Goel et al. (2025) reported a family enhancement bias where models favor outputs from the same model family.

Input Sensitivity in LLM Judges A complementary line studies LLM sensitivity to input variations. Wei et al. (2024) identify order/token sensitivity in multiple-choice tasks and mitigate it via dataset-level probability weighting and calibration. We instead study score range bias in direct assessment, where shifting an equivalent 5-point range (e.g., 0-4 vs. 2-6) changes judge-to-human correlation; we further show this bias is shared across same-family models (§4.2) and mitigate it via instance-level contrastive decoding that cancels family-shared bias.

3 Analysis and Mitigation

3.1 Identifying Score Range Bias

An ideal LLM judge should maintain consistent correlation with human judgments across shifted ranges, as they represent the same 5-point scale (e.g., 0-4, 1-5, 2-6, 3-7). Failure to do so indicates **score range bias**: models exhibit different correlations depending on the score range used, even when evaluating identical content. This bias outputs skewed distributions where models, for example, favor specific scores. We hypothesize that these biases can be mitigated with contrastive decoding, which cancels out shared biases.

3.2 Mitigation by Contrastive Decoding

Contrastive decoding (Li et al., 2023) modifies the model outputs by using two models: a main model and an assistant model. Given the next token probability of a main model p_{main} and an assistant model p_{asst}, the final adjusted score is calculated by subtracting the weighted p_{asst} from p_{main} i.e.,

$$\log p_{\text{main}} - \lambda \log p_{\text{asst}} \tag{1}$$

where $\lambda \in \mathbb{R}$ is the hyperparameter to control the magnitude of assistant model and logit e_i of token i is controlled by temperature $t > 0$ i.e., $p_{\text{Asst}} = \frac{e_i/t}{\sum_j e_j/t}$. We depart from Li et al. (2023) by introducing λ because logit magnitudes differ across sizes (e.g., max logit ≈ 25 for Qwen-2.5-3B vs. ≈ 34 for 14B) and this mismatch carries over to log-probabilities; without λ the assistant can undercorrect or erase the main model's signal. We tune λ on a small held-out set (Appendix B).

4 Experiments

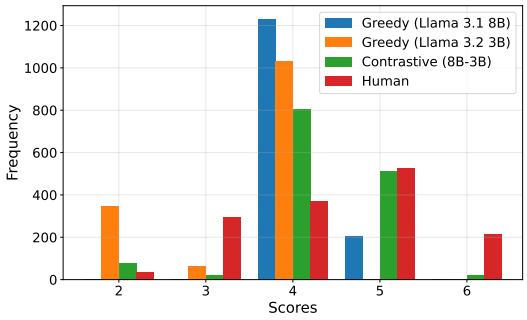
We focus on direct assessment on summarization since prior work reported that LLM judges fall short (Ye et al., 2024) and they are commonly evaluated on summarization (Panickssery et al., 2024).

4.1 Setup

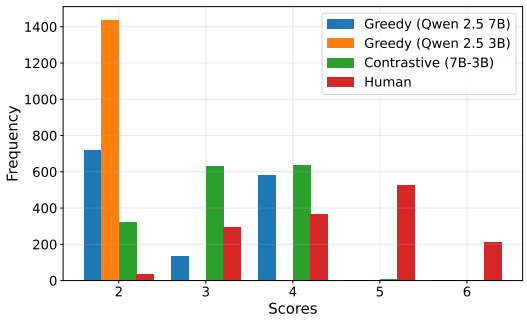
Task and Metrics We focus on the summarization task where LLM judges are commonly used (Liu et al., 2023; Panickssery et al., 2024). The correlations between human annotations are measured using three metrics: Pearson, Spearman, and Kendall correlations.

Score Scale and Ranges We use the 5 points Likert scale (Likert, 1932) on different score ranges (0-4, 1-5, 2-6, 3-7)[2]. If output score parsing fails, we set to the lowest score following Liu et al. (2023) and if the parsed score exceeds the maximum, we clamp to the highest score in the range.

[2]We stopped at 7 inspired by Likert (1932) showing high correlation between 5 points (1-5) and 7 points (1-7) results.



(a) Llama-3 Family Results



(b) Qwen2.5 Family Results

Figure 2: Coherence score distribution in 2-6 score range with greedy decoding, contrastive decoding, and human annotations. The greedy decoding outputs from both Llama 8B and 3B models are highly skewed towards outputting score of 4. Qwen2.5 3B (see also Figure 3) and 7B models are outputting score of 2 showing similar biases are encoded in these models.

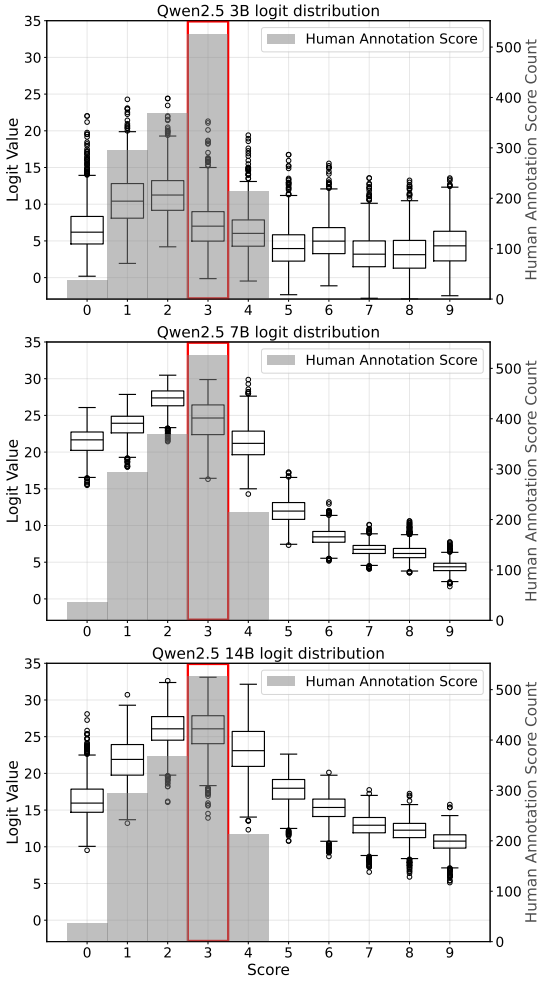


Figure 3: Logit distribution of the first output token in the 0-4 score range for Qwen2.5 3B, 7B, and 14B, with human annotation frequencies overlaid (gray bars). As model size increases from 3B to 14B, Score 3's logit (red rectangle) grows and moves closer to the human annotations.

Models We experiment on two model families.[3] For Llama-3 family (Grattafiori et al., 2024), we use `Llama-3.1-8B-Instruct` as the main model, `Llama-3.2-3B-Instruct` and `Llama-3.2-1B-Instruct` as the assistant model,[4] and for Qwen2.5 family (Qwen et al., 2025), we use `Qwen-2.5-14B-Instruct` and `Qwen-2.5-7B-Instruct` as the main models, and `Qwen-2.5-3B-Instruct` as the assistant model. See Appendix A for the prompt used.

Dataset We use SummEval (Fabbri et al., 2021), a summarization benchmark also used by Liu et al. (2023) which contains 100 news articles where each article is associated with 16 summaries with human annotation scores, which sums up to 1600 summaries. 10% of the news articles are used as the held out development set to conduct grid search.

We now first reveal the score range bias of LLM judges by evaluating correlation to human annotations in the same 5 points scale but with different score ranges, and then experiment on mitigating it.

4.2 Reveal and Mitigate Score Range Biases

Similar score range biases exist across models from the same family We first analyze the distribution of the output scores in the 2 to 6 score range (Figure 2). Llama family models (3B and 8B) tend to output score of 4 (Figure 2a) and Qwen 2.5 family models tend to output score of 2 (Figure 2b). By using contrastive decoding, these biases toward

[3]We leave models like Prometheus (Kim et al., 2024) specifically fineutuned on judge tasks as future work since multiple model sizes are not available for contrastive decoding and those models are finetuned towards 1-5 score range.

[4]Although Llama-3.1 and Llama-3.2 are versioned separately, Meta documents them as part of the same Llama 3 herd (Grattafiori et al., 2024) with a shared tokenizer and architecture family. Figure 2 empirically supports it: Llama-3.1-8B and Llama-3.2-3B exhibit the same score range bias, both skewing toward Score 4 in the 2-6 range.

Model	Range	Pear.	Spear.	Kend.
Llama 3.2-1B	0 to 4	$.073_{.036}$	$.055_{.035}$	$.047_{.016}$
Llama 3.2-3B	0 to 4	$.056_{.033}$	$.089_{.028}$	$.073_{.026}$
Llama 3.1-8B	0 to 4	$\underline{.407_{.029}}$	$\underline{.384_{.036}}$	$\underline{.327_{.022}}$
Contra. (8B-1B)	0 to 4	$.384_{.037}$	$.363_{.038}$	$.309_{.025}$
Contra. (8B-3B)	0 to 4	$.380_{.030}$	$.357_{.048}$	$.303_{.040}$
Llama 3.2-1B	1 to 5	$.000_{.000}$	$.000_{.000}$	$.000_{.000}$
Llama 3.2-3B	1 to 5	$.151_{.036}$	$.201_{.042}$	$.163_{.016}$
Llama 3.1-8B	1 to 5	$.338_{.036}$	$.309_{.042}$	$.262_{.029}$
Contra. (8B-1B)	1 to 5	$\underline{.357_{.021}}$	$\underline{.390_{.049}}$	$\underline{.327_{.030}}$
Contra. (8B-3B)	1 to 5	$.323_{.043}$	$.315_{.029}$	$.265_{.029}$
Llama 3.2-1B	2 to 6	$.000_{.000}$	$.000_{.000}$	$.000_{.000}$
Llama 3.2-3B	2 to 6	$.008_{.032}$	$.031_{.029}$	$.026_{.027}$
Llama 3.1-8B	2 to 6	$.262_{.029}$	$.257_{.025}$	$.220_{.015}$
Contra. (8B-1B)	2 to 6	$\underline{.344_{.038}}$	$\underline{.337_{.035}}$	$\underline{.288_{.034}}$
Contra. (8B-3B)	2 to 6	$.302_{.049}$	$.305_{.034}$	$.254_{.023}$
Llama 3.2-1B	3 to 7	$-.039_{.035}$	$-.051_{.020}$	$-.043_{.028}$
Llama 3.2-3B	3 to 7	$-.007_{.035}$	$-.009_{.044}$	$-.008_{.027}$
Llama 3.1-8B	3 to 7	$\underline{.445_{.027}}$	$\underline{.426_{.030}}$	$\underline{.352_{.030}}$
Contra. (8B-1B)	3 to 7	$.425_{.030}$	$.408_{.033}$	$.341_{.019}$
Contra. (8B-3B)	3 to 7	$.442_{.018}$	$.419_{.035}$	$.347_{.023}$
Average across all score ranges				
Llama 3.2-1B		.009	.001	.001
Llama 3.2-3B		.052	.078	.064
Llama 3.1-8B		.363	.344	.290
Contra. (8B-1B)		**.378**	**.375**	**.316**
Contra. (8B-3B)		.362	.349	.292

Table 1: Llama-3 family correlation to humans on summary coherence with 95% confidence interval from bootstrap testing. Max correlation within score range are underlined and max averages are **bolded**. Maximal improvement is observed in the 2-6 score range.

Model	Range	Pear.	Spear.	Kend.
Qwen2.5-3B	0 to 4	$-.042_{.038}$	$-.059_{.023}$	$-.051_{.028}$
Qwen2.5-7B	0 to 4	$.248_{.048}$	$.245_{.028}$	$.209_{.018}$
Qwen2.5-14B	0 to 4	$.435_{.041}$	$.452_{.019}$	$.382_{.021}$
Contra. (7B-3B)	0 to 4	$.334_{.033}$	$.332_{.048}$	$.282_{.023}$
Contra. (14B-3B)	0 to 4	$\underline{.442_{.037}}$	$\underline{.458_{.031}}$	$\underline{.385_{.025}}$
Qwen2.5-3B	1 to 5	$-.044_{.042}$	$-.056_{.040}$	$-.048_{.015}$
Qwen2.5-7B	1 to 5	$.382_{.043}$	$.370_{.042}$	$.309_{.024}$
Qwen2.5-14B	1 to 5	$\underline{.468_{.040}}$	$.468_{.026}$	$.385_{.025}$
Contra. (7B-3B)	1 to 5	$.365_{.042}$	$.356_{.056}$	$.297_{.019}$
Contra. (14B-3B)	1 to 5	$.463_{.039}$	$\underline{.470_{.041}}$	$\underline{.387_{.027}}$
Qwen2.5-3B	2 to 6	$-.010_{.000}$	$-.013_{.000}$	$-.011_{.000}$
Qwen2.5-7B	2 to 6	$.362_{.024}$	$.355_{.036}$	$.293_{.025}$
Qwen2.5-14B	2 to 6	$.310_{.037}$	$.313_{.056}$	$.260_{.026}$
Contra. (7B-3B)	2 to 6	$.373_{.041}$	$.364_{.034}$	$.298_{.021}$
Contra. (14B-3B)	2 to 6	$\underline{.410_{.028}}$	$\underline{.426_{.035}}$	$\underline{.359_{.040}}$
Qwen2.5-3B	3 to 7	$.033_{.016}$	$.034_{.000}$	$.029_{.012}$
Qwen2.5-7B	3 to 7	$.341_{.045}$	$.349_{.024}$	$.289_{.037}$
Qwen2.5-14B	3 to 7	$.349_{.039}$	$.336_{.036}$	$.275_{.036}$
Contra. (7B-3B)	3 to 7	$.351_{.019}$	$.357_{.034}$	$.298_{.027}$
Contra. (14B-3B)	3 to 7	$\underline{.413_{.027}}$	$\underline{.398_{.020}}$	$\underline{.327_{.029}}$
Average across all score ranges				
Qwen2.5-3B		-.016	-.024	-.020
Qwen2.5-7B		.333	.330	.275
Qwen2.5-14B		.391	.392	.326
Contra. (7B-3B)		.356	.352	.294
Contra. (14B-3B)		**.432**	**.438**	**.365**

Table 2: Qwen-2.5 family correlation to humans on summary coherence with 95% confidence interval from bootstrap testing. Max correlation within score range are underlined and max averages are **bolded**. Maximal improvement is observed in the 2-6 score range.

specific ranges are mitigated and making the score outputs closer to human annotations.

Upon analyzing the first output token logit distribution (Figure 3) of the Qwen family models in the 0-4 score range, Qwen-2.5 3B, 7B, 8B, and 14B models encode similar biases where Score 2 is the highest while the most frequent human annotation is Score 3. The bias towards Score 2 gradually decreases as the model size scales from 3B to 14B, but still remains even in the 14B model. Furthermore, the logit range in each model differs e.g., max logit in 3B $\approx$ 25, 7B $\approx$ 30, and 14B $\approx$ 34 (Figure 3), motivating the inclusion of λ in Eq. 1 to align the distributions between these models. This bias on Score 2 in the 3B model helps decrease similar bias encoded in the 7B and 14B models when used as the assistant model.

As a result of score range bias, using Llama 3B or 7B with greedy decoding causes the lowest correlation in 2-6 score range (Table 1). This trend is not limited to the Llama-3 family models and it is also observed in the Qwen-2.5 family models (Table 2). Focusing on greedy decoding, Qwen-2.5 family and Llama3.1-3B show clearer trend that the 1-5 score range shows the highest correlation among the experimented score ranges (7B, 1-5: .370, 14B, 1-5: .468), while Llama3.1-8B being the exception that 3 to 7 score range showing highest correlation among all score ranges. These outcomes further raises concern on applying LLM judges beyond the standard 1-5 range.

Contrastive Decoding is a Robust Mitigation Strategy across Different Score Ranges We use *robust* here to mean reducing variance in judge-to-human correlation across score ranges, rather than uniformly improving at every individual range. Table 1 and 2 show that contrastive decoding exhibits this robustness: correlations remain more consistent across varying score ranges, directly addressing the score range bias observed. While using a single model suffers from decrease in correlation when score ranges are shifted, contrastive decoding maintains more stable correlations with human judgments regardless of the score ranges (Table 1). This robustness is evident in the 2-6 range, where

contrastive decoding on Llama-3 family achieves a Pearson correlation of .310 (compared to .168 for Llama 3.2-3B and .270 for Llama 3.1-8B) and similar improvements in Spearman and Kendall correlations, also seen as 6.7% relative improvement for Llama 8B (.330 → .352) and 11.7% for Qwen 14B (.392 → .438) on average across all score ranges on SummEval. The stability across different scoring ranges enables search on optimal score ranges beyond the 1-5 range (e.g., 0-4 range showing the best correlation in summary relevance for Qwen family in Appendix D).

Does Assistant Model Choice Matter for Bias Mitigation? Table 1 shows that the choice of assistant model slightly impacts correlations, with the 1B model marginally outperforming the 3B model. The 1B assistant achieves an average Spearman correlation of .375 compared to .352 for the 3B assistant. Notably, using larger assistant models can degrade performance: our ablation study with Qwen 14B-7B (Appendix H) shows significant degradation in the 1-5 range (Spearman: .282 vs .470 with 3B assistant), showing that larger assistants penalize correct logits from the main model.

Generalization to BigGen-Bench To validate that our findings extend beyond summarization, we replicate our analysis on BigGen-Bench (Kim et al., 2025), a fine-grained LLM evaluation benchmark spanning 77 tasks covering categories such as Reasoning, Dialogue/Generation, Safety, Planning, Tool Usage, Instruction Following, Theory of Mind, and Grounding, each paired with task-specific rubrics and human gold scores. We apply the same judge setup as in our SummEval experiments: the 5-point Likert scale with the four score ranges (0-4, 1-5, 2-6, 3-7), identical prompting, and the benchmark's human gold scores as the correlation reference. Tables 3 and 4 show that contrastive decoding improves the average Spearman correlation with human judgments from .349 to .352 for Llama 3.1-8B (1B assistant) and from .500 to .517 for Qwen2.5-14B (3B assistant), consistent with the trends observed on SummEval.

Model	Range	Pear.	Spear.	Kend.
Llama 3.1-8B	0 to 4	$.384_{.035}$	$.361_{.034}$	$.301_{.029}$
Contra. (8B-1B)	0 to 4	$\underline{.395_{.036}}$	$\underline{.367_{.034}}$	$\underline{.307_{.029}}$
Llama 3.1-8B	1 to 5	$.377_{.035}$	$.345_{.035}$	$.289_{.030}$
Contra. (8B-1B)	1 to 5	$\underline{.392_{.036}}$	$\underline{.352_{.034}}$	$\underline{.300_{.030}}$
Llama 3.1-8B	2 to 6	$.353_{.037}$	$.336_{.035}$	$.284_{.030}$
Contra. (8B-1B)	2 to 6	$\underline{.359_{.035}}$	$\underline{.349_{.034}}$	$\underline{.294_{.029}}$
Llama 3.1-8B	3 to 7	$\underline{.373_{.034}}$	$\underline{.354_{.034}}$	$\underline{.300_{.030}}$
Contra. (8B-1B)	3 to 7	$.357_{.037}$	$.338_{.036}$	$.288_{.031}$
Average across all score ranges				
Llama 3.1-8B		.372	.349	.294
Contra. (8B-1B)		**.376**	**.352**	**.297**

Table 3: Llama 3.1 family correlation results on BigGen-Bench with 95% bootstrap confidence intervals. Max correlation within score range are underlined and max averages are **bolded**.

Model	Range	Pear.	Spear.	Kend.
Qwen2.5-14B	0 to 4	$.545_{.031}$	$.517_{.030}$	$.445_{.027}$
Contra. (14B-3B)	0 to 4	$\underline{.571_{.029}}$	$\underline{.533_{.030}}$	$\underline{.460_{.026}}$
Qwen2.5-14B	1 to 5	$.520_{.033}$	$.489_{.032}$	$.427_{.028}$
Contra. (14B-3B)	1 to 5	$\underline{.534_{.033}}$	$\underline{.498_{.032}}$	$\underline{.435_{.028}}$
Qwen2.5-14B	2 to 6	$.533_{.032}$	$.501_{.032}$	$.436_{.028}$
Contra. (14B-3B)	2 to 6	$\underline{.556_{.031}}$	$\underline{.530_{.031}}$	$\underline{.459_{.027}}$
Qwen2.5-14B	3 to 7	$.528_{.033}$	$.494_{.033}$	$.432_{.029}$
Contra. (14B-3B)	3 to 7	$\underline{.539_{.032}}$	$\underline{.507_{.030}}$	$\underline{.443_{.028}}$
Average across all score ranges				
Qwen2.5-14B		.532	.500	.435
Contra. (14B-3B)		**.550**	**.517**	**.449**

Table 4: Qwen2.5 family correlation results on BigGen-Bench with 95% bootstrap confidence intervals. Max correlation within score range are underlined, and max averages are **bolded**. Qwen2.5 14B with contrastive decoding (14B-3B) consistently achieves the best performance across all score ranges.

5 Conclusion

In this work, we analyze and experiment with LLM-as-a-judge on direct assessment, which reveals two key findings: First, LLM judges exhibit a score range bias across different model families and sizes with a tendency to favor specific scores regardless of the quality of the summaries. Second, we show that contrastive decoding effectively mitigates score range bias by leveraging the similar biases present in models from the same family. Our score range bias analysis framework can test arbitrary models and can help unlock the potential to expand beyond the standard 1-5 score range.

Limitations

Inference Time Compute Contrastive learning increases the test time compute due to running forward pass on two models rather than one. On the other hand, using a main model and an assistant model is very common in real world setup to speed up decoding with speculative decoding (Leviathan et al., 2023), and therefore contrastive decoding can be used without an additional forward pass when

speculative decoding is used.

Model Size Our experiments were limited to models with up to 14B parameters due to computational budget constraints.

Language Coverage Our experiments are conducted only on English language, however, we have not exploited linguistic knowledge specific to English.

Acknowledgments

We sincerely thank the anonymous reviewers for their constructive feedback and insightful comments, which helped strengthen the analysis and improve the clarity of this work.

References

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *Preprint*, arXiv:2303.12712.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Andrew Gambardella, Yusuke Iwasawa, and Yutaka Matsuo. 2024. Language models do hard arithmetic tasks easily and hardly do easy arithmetic tasks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 85–91, Bangkok, Thailand. Association for Computational Linguistics.

Shashwat Goel, Joschka Struber, Ilze Amanda Auzina, Karuna K Chandra, Ponnurangam Kumaraguru, Douwe Kiela, Ameya Prabhu, Matthias Bethge, and Jonas Geiping. 2025. Great models think alike and this undermines ai oversight. *Preprint*, arXiv:2502.04313.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xin-

feng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Jaylen Jones, Lingbo Mo, Eric Fosler-Lussier, and Huan Sun. 2024. A multi-aspect framework for counter narrative evaluation using large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 147–168, Mexico City, Mexico. Association for Computational Linguistics.

Seungone Kim, Juyoung Suk, Ji Yong Cho, Shayne Longpre, Chaeeun Kim, Dongkeun Yoon, Guijin Son, Yejin Cho, Sheikh Shafayat, Jinheon Baek, Sue Hyun Park, Hyeonbin Hwang, Jinkyung Jo, Hyowon Cho, Haebin Ravi, Hadi Pouransari, Suyeon Kevin Ha, Cheng Chen, Jiwoo Roh, Chirag Raman, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2025. The BIGGEN BENCH: A principled benchmark for fine-grained evaluation of language models with language models. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics*.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024. Prometheus 2: An open source language model specialized in evaluating other language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4334–4353, Miami, Florida, USA. Association for Computational Linguistics.

Yaniv Leviathan, Matan Kalman, and Yossi Matias. 2023. Fast inference from transformers via speculative decoding. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Junlong Li, Shichao Sun, Weizhe Yuan, Run-Ze Fan, hai zhao, and Pengfei Liu. 2024. Generative judge for evaluating alignment. In *The Twelfth International Conference on Learning Representations*.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2023. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12286–12312, Toronto, Canada. Association for Computational Linguistics.

Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of psychology*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Yinhong Liu, Han Zhou, Zhijiang Guo, Ehsan Shareghi, Ivan Vulić, Anna Korhonen, and Nigel Collier. 2024. Aligning with human judgement: The role of pairwise preference in large language model evaluators. In *First Conference on Language Modeling*.

Rodrigo Nogueira, Zhiying Jiang, and Jimmy Lin. 2021. Investigating the limitations of transformers with simple arithmetic tasks. *Preprint*, arXiv:2102.13019.

Sean O'Brien and Mike Lewis. 2023. Contrastive decoding improves reasoning in large language models. *Preprint*, arXiv:2309.09117.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. 2024. LLM evaluators recognize and favor their own generations. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Koki Wataoka, Tsubasa Takahashi, and Ryokan Ri. 2024. Self-preference bias in LLM-as-a-judge. In *Neurips Safe Generative AI Workshop 2024*.

Sheng-Lun Wei, Cheng-Kuang Wu, Hen-Hsen Huang, and Hsin-Hsi Chen. 2024. Unveiling selection biases: Exploring order and token sensitivity in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5598–5621, Bangkok, Thailand. Association for Computational Linguistics.

Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, Nitesh V Chawla, and Xiangliang Zhang. 2024. Justice or prejudice? quantifying biases in llm-as-a-judge. *Preprint*, arXiv:2410.02736.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Lianghui Zhu, Xinggang Wang, and Xinlong Wang. 2025. JudgeLM: Fine-tuned large language models are scalable judges. In *The Thirteenth International Conference on Learning Representations*.

A Judge Prompts

We use the following prompt experimented by Liu et al. (2023).

Score Range {min_range}-{max_range} for Coherence

You will be given one summary written for a news article.

Your task is to rate the summary on one metric.

Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed.

Evaluation Criteria:

Coherence ({min_range}-{max_range}) - the collective quality of all sentences. We align this dimension with the DUC quality question of structure and coherence whereby "the summary should be well-structured and well-organized. The summary should not just be a heap of related information, but should build from sentence to a coherent body of information about a topic."

Evaluation Steps:

1. Read the news article carefully and identify the main topic and key points.
2. Read the summary and compare it to the news article. Check if the summary covers the main topic and key points of the news article, and if it presents them in a clear and logical order.
3. Assign a score for coherence on a scale of {min_range} to {max_range}, where {min_range} is the lowest and {max_range} is the highest based on the Evaluation Criteria.

Example:

Source Text:

{{Document]}}

Summary:

{{Summary}}

Evaluation Form (scores ONLY):

- Coherence:

What is the coherence of the summary above? Provide only rating and no other text.

Score Range {min_range}-{max_range} for Relevance

You will be given one summary written for a news article.

Your task is to rate the summary on one metric.

Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed.

Evaluation Criteria:

Relevance ({min_range}-{max_range}) - selection of important content from the source. The summary should include only important information from the source document. Annotators were instructed to penalize summaries which contained redundancies and excess information.

Evaluation Steps:

1. Read the summary and the source document carefully.
2. Compare the summary to the source document and identify the main points of the article.
3. Assess how well the summary covers the main points of the article, and how much irrelevant or redundant information it contains.
4. Assign a relevance score from {min_range} to {max_range}.

Example:

Source Text:

{{Document]}}

Summary:

{{Summary}}

Evaluation Form (scores ONLY):

- Relevance:

What is the relevance of the summary above? Provide only rating and no other text.

Score Range {min_range}-{max_range} for Consistency

You will be given one summary written for a news article.

Your task is to rate the summary on one metric.

Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed.

Evaluation Criteria:

Consistency ({min_range}-{max_range}) - the factual alignment between the summary and the summarized source. A factually consistent summary contains only statements that are entailed by the source document. Annotators were also asked to penalize summaries that contained hallucinated facts.

Evaluation Steps:

1. Read the news article carefully and identify the main topic and key points.
2. Read the summary and compare it to the news article. Check if the summary covers the main topic and key points of the news article, and if it presents them in a clear and logical order.
3. Assign a score for consistency based on the Evaluation Criteria.

Example:

Source Text:

{{Document]}}

Summary:

{{Summary}}

Evaluation Form (scores ONLY):

- Consistency:

What is the consistency of the summary above? Provide only rating and no other text.

B Hyper-parameters

We conduct grid search over two hyperparameters for contrastive decoding: 1) temperature t and 2) scaling constant λ from the following ranges:

- $\lambda = [0.01, 0.1, 0.5, 1.0]$
- $t = [0.5, 1.0, 2.0, 3.0, 4.0, 5.0]$

The following table shows the hyperparameters setup for each setting:

Main	Asst	Range	λ	t
Llama 3.1 8B	Llama 3.2 3B	0-4	0.01	1.0
		1-5	1.0	0.5
		2-6	1.0	0.5
		3-7	0.01	5.0
	Llama 3.2 1B	0-4	0.01	0.5
		1-5	0.1	5.0
		2-6	0.1	2.0
		3-7	0.1	2.0
Qwen 2.5 7B	Qwen 2.5 3B	0-4	0.1	4.0
		1-5	0.01	5.0
		2-6	0.1	4.0
		3-7	0.01	0.5
Qwen 2.5 14B	Qwen 2.5 3B	0-4	0.1	2.0
		1-5	0.01	4.0
		2-6	0.1	1.0
		3-7	0.1	2.0

Table 5: Hyperparameter settings for contrastive decoding for each main and assistant model pair from each model family for evaluating summary coherence.

Main	Asst	Range	λ	t
Llama 3.1 8B	Llama 3.2 3B	0-4	0.01	0.5
		1-5	0.01	0.5
		2-6	0.01	0.5
		3-7	0.5	0.5
	Llama 3.2 1B	0-4	0.01	0.5
		1-5	0.1	5.0
		2-6	0.1	5.0
		3-7	0.01	0.5
Qwen 2.5 7B	Qwen 2.5 3B	0-4	0.1	5.0
		1-5	0.5	1.0
		2-6	1.0	1.0
		3-7	0.1	4.0
Qwen 2.5 14B	Qwen 2.5 3B	0-4	0.01	3.0
		1-5	0.1	0.5
		2-6	0.01	3.0
		3-7	0.01	3.0

Table 6: Hyperparameter settings for contrastive decoding for each Llama-3 main and assistant model pair for relevance.

C Score Distribution Across All Ranges

Figures 4, 5, and 6 show the score distribution comparison plots across all four score ranges (0-4, 1-5, 2-6, 3-7) for Llama-3, Qwen2.5 7B, and Qwen2.5 14B family models respectively. These plots show how score range bias manifests differently across ranges and model sizes, and how contrastive decoding consistently mitigates this bias.

Main	Asst	Range	λ	t
Llama 3.1 8B	Llama 3.2 3B	0-4	0.01	5.0
		1-5	0.1	2.0
		2-6	0.1	2.0
		3-7	0.1	3.0
	Llama 3.2 1B	0-4	0.1	1.0
		1-5	0.1	0.5
		2-6	0.1	2.0
		3-7	0.1	1.0
Qwen 2.5 7B	Qwen 2.5 3B	0-4	0.1	3.0
		1-5	0.01	5.0
		2-6	0.01	2.0
		3-7	0.01	5.0
Qwen 2.5 14B	Qwen 2.5 3B	0-4	0.1	1.0
		1-5	0.1	5.0
		2-6	0.1	3.0
		3-7	0.1	3.0

Table 7: Hyperparameter settings for contrastive decoding for each Llama-3 main and assistant model pair for consistency.

Model	Range	Invalid	Failure Types
Llama 3.1 8B	0-4	0.0%	-
	1-5	0.0%	-
	2-6	0.0%	-
	3-7	0.0%	-
Llama 3.2 3B	0-4	23.2%	No Pred: 334
	1-5	17.4%	No Pred: 251
	2-6	24.0%	No Pred: 345
	3-7	18.6%	No Pred: 268
Llama 3.1 8B Contrastive	0-4	0.0%	-
	1-5	0.0%	-
	2-6	0.3%	No Pred: 5
	3-7	0.0%	-

Table 8: Invalid output statistics for Llama-3 models across all score ranges on coherence evaluation. The 8B model shows perfect parsing, while the 3B model exhibits moderate failure rates of 17-24%.

C.1 Invalid Output Analysis

Tables 8 and 9 show the analysis of invalid model outputs across different score ranges for Llama-3 and Qwen2.5 models. Invalid outputs are classified into three categories: (1) *No Pred*: outputs where no numeric pattern could be extracted, and (2) *Below Range*: outputs with valid numbers below the minimum score for the range.

For Llama models, the 8B model achieves near-perfect parsing with 0% invalid outputs across all ranges. The 3B model shows moderate failure rates ranging from 17.4% to 24.0%, primarily due to malformed outputs. Notably, contrastive decoding maintains excellent performance with minimal failures (0.3% in the 2-6 range).

For Qwen models, the analysis reveals that the 3B model exhibits catastrophic failure with approximately 98% invalid outputs across all score ranges, primarily due to malformed outputs that cannot be parsed. In contrast, the 7B and 14B models show significantly better performance, with the 7B model achieving near-perfect parsing (0% invalid) and the 14B model showing minimal failures, primarily in the 3-7 range where 4.6% of outputs fall below the minimum score.

D Relevance and Consistency Results

In Tables 10 and 11, we present the correlation results for summary relevance evaluation across different score ranges for the Llama-3 and Qwen2.5 model families, respectively. In Tables 12 and 13, we present the correlation results for summary consistency evaluation across different score ranges.

E Model Size and Budget

For all the experiments in this paper, NVIDIA's A100 GPU was used. The base models used in this paper are licensed under Meta Llama 3 License[5] for Llama 3 family models and Apache-2.0 license for Qwen 2.5 family models. We followed their intended use case.

F Information About Use of AI Assistants

We have used Claude on this manuscript to enhance the clarity of the paper and fixing grammatical mistakes. We also used it to create the codes to run experiments.

G Potential Risks

As discussed in the limitations section, the experiments are only conducted on English, which may bias the takeaways on English.

H Assistant Model Size Analysis

The paper notes that smaller assistant models sometimes work better than larger ones, but does not explore why. We conducted a follow-up experiment using Qwen2.5-7B as an assistant model (instead of the 3B model) for the Qwen2.5-14B main model on summary coherence evaluation.

Table 14 shows the results for Qwen 14B-7B contrastive decoding across all score ranges. We observe no significant changes in the 2-6 and 3-7 ranges compared to using the 3B assistant. However, there is a notable degradation in the 1-5 range, where the 7B assistant model significantly reduces

[5] https://www.llama.com/llama3/license/

Model	Range	Invalid	Failure Types
Qwen2.5 3B	0-4	98.2%	No Pred: 1414
	1-5	98.7%	No Pred: 1421
	2-6	97.6%	No Pred: 1406
	3-7	97.6%	No Pred: 1376, B: 30
Qwen2.5 7B	0-4	0.0%	-
	1-5	0.0%	-
	2-6	0.0%	-
	3-7	0.0%	-
Qwen2.5 14B	0-4	0.0%	-
	1-5	0.0%	-
	2-6	0.1%	B: 1
	3-7	4.6%	B: 66
Qwen2.5 7B Contrastive	0-4	0.0%	-
	1-5	0.0%	-
	2-6	0.0%	-
	3-7	0.0%	-
Qwen2.5 14B Contrastive	0-4	0.0%	-
	1-5	0.0%	-
	2-6	0.1%	B: 1
	3-7	5.6%	B: 81

Table 9: Invalid output statistics for Qwen2.5 models across all score ranges on coherence evaluation. The 3B model shows near-complete failure in generating valid scores, while larger models perform significantly better. B: Below

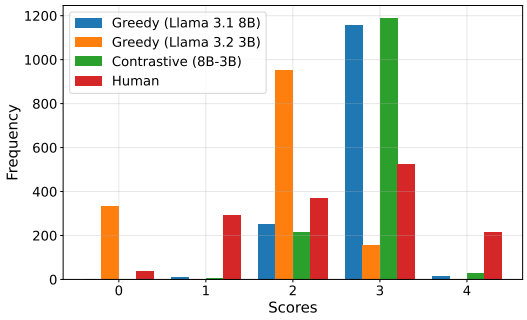
correlation (Spearman: 0.282 vs 0.470 with 3B assistant, see Table 2). This finding supports the hypothesis that using a larger assistant model is more likely to penalize correct logits from the main model, thereby degrading performance.

Model	Range	Pear.	Spear.	Kend.
Llama 3.2-1B	0 to 4	$.028_{.041}$	$.028_{.048}$	$.024_{.037}$
Llama 3.2-3B	0 to 4	$.072_{.027}$	$.111_{.036}$	$.094_{.025}$
Llama 3.1-8B	0 to 4	$\underline{.429_{.050}}$	$\underline{.374_{.027}}$	$\underline{.323_{.016}}$
Contra. (8B-3B)	0 to 4	$.407_{.036}$	$.360_{.038}$	$.310_{.027}$
Llama 3.2-1B	1 to 5	$.014_{.054}$	$.022_{.036}$	$.019_{.033}$
Llama 3.2-3B	1 to 5	$.161_{.056}$	$.179_{.023}$	$.152_{.023}$
Llama 3.1-8B	1 to 5	$\underline{.391_{.042}}$	$\underline{.341_{.025}}$	$\underline{.293_{.019}}$
Contra. (8B-3B)	1 to 5	$.375_{.030}$	$.322_{.035}$	$.277_{.023}$
Llama 3.2-1B	2 to 6	$.000_{.000}$	$.000_{.000}$	$.000_{.000}$
Llama 3.2-3B	2 to 6	$.076_{.041}$	$.107_{.037}$	$.091_{.041}$
Llama 3.1-8B	2 to 6	$\underline{.378_{.020}}$	$\underline{.360_{.033}}$	$\underline{.312_{.032}}$
Contra. (8B-3B)	2 to 6	$.393_{.033}$	$.381_{.030}$	$.331_{.025}$
Llama 3.2-1B	3 to 7	$.036_{.018}$	$.033_{.018}$	$.029_{.016}$
Llama 3.2-3B	3 to 7	$.092_{.044}$	$.095_{.040}$	$.082_{.037}$
Llama 3.1-8B	3 to 7	$.438_{.041}$	$.421_{.021}$	$.357_{.019}$
Contra. (8B-3B)	3 to 7	$\underline{.471_{.035}}$	$\underline{.444_{.023}}$	$\underline{.369_{.028}}$
Average across all score ranges				
Llama 3.2-1B		.026	.028	.024
Llama 3.2-3B		.100	.123	.105
Llama 3.1-8B		.409	.374	.321
Contra. (8B-3B)		**.412**	**.377**	**.322**

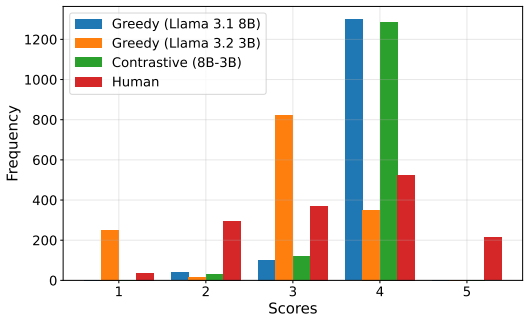
Table 10: Llama-3 family correlation results to human annotations on summary relevance. Max correlation within score range are underlined, max across all ranges are *italicized*, and max averages are **bolded**.

Model	Range	Pear.	Spear.	Kend.
Qwen2.5-3B	0 to 4	$.000_{.000}$	$.000_{.000}$	$.000_{.000}$
Qwen2.5-7B	0 to 4	$.323_{.028}$	$.296_{.041}$	$.250_{.045}$
Qwen2.5-14B	0 to 4	$\underline{.517_{.033}}$	$\underline{.496_{.034}}$	$\underline{.416_{.023}}$
Contra. (7B-3B)	0 to 4	$.351_{.031}$	$.289_{.037}$	$.236_{.053}$
Contra. (14B-3B)	0 to 4	$.516_{.031}$	$.478_{.038}$	$.406_{.016}$
Qwen2.5-3B	1 to 5	$.000_{.000}$	$.000_{.000}$	$.000_{.000}$
Qwen2.5-7B	1 to 5	$.345_{.037}$	$.347_{.043}$	$.285_{.021}$
Qwen2.5-14B	1 to 5	$\underline{.518_{.036}}$	$\underline{.500_{.036}}$	$\underline{.427_{.033}}$
Contra. (7B-3B)	1 to 5	$.353_{.045}$	$.335_{.036}$	$.283_{.043}$
Contra. (14B-3B)	1 to 5	$.509_{.024}$	$.495_{.034}$	$.422_{.024}$
Qwen2.5-3B	2 to 6	$.000_{.000}$	$.000_{.000}$	$.000_{.000}$
Qwen2.5-7B	2 to 6	$.415_{.040}$	$.398_{.024}$	$.330_{.027}$
Qwen2.5-14B	2 to 6	$.438_{.034}$	$.380_{.038}$	$.319_{.030}$
Contra. (7B-3B)	2 to 6	$.171_{.020}$	$.134_{.024}$	$.112_{.027}$
Contra. (14B-3B)	2 to 6	$\underline{.453_{.036}}$	$\underline{.408_{.027}}$	$\underline{.344_{.024}}$
Qwen2.5-3B	3 to 7	$.000_{.000}$	$.000_{.000}$	$.000_{.000}$
Qwen2.5-7B	3 to 7	$.468_{.023}$	$.441_{.035}$	$.367_{.020}$
Qwen2.5-14B	3 to 7	$.426_{.034}$	$.375_{.037}$	$.313_{.023}$
Contra. (7B-3B)	3 to 7	$.456_{.035}$	$.438_{.024}$	$.365_{.024}$
Contra. (14B-3B)	3 to 7	$\underline{.510_{.026}}$	$\underline{.473_{.038}}$	$\underline{.394_{.016}}$
Average across all score ranges				
Qwen2.5-3B		.000	.000	.000
Qwen2.5-7B		.388	.371	.308
Qwen2.5-14B		.475	.438	.369
Contra. (7B-3B)		.333	.299	.249
Contra. (14B-3B)		**.497**	**.464**	**.392**

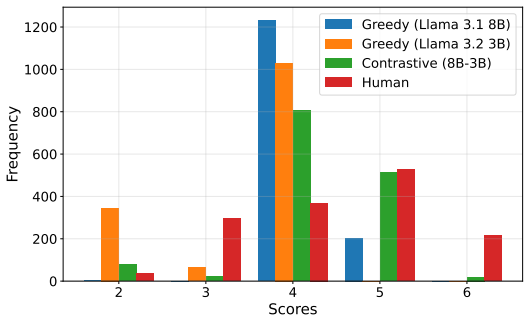
Table 11: Qwen2.5 family correlation results to human annotations on summary relevance. Max correlation within score range are underlined and max averages are **bolded**.



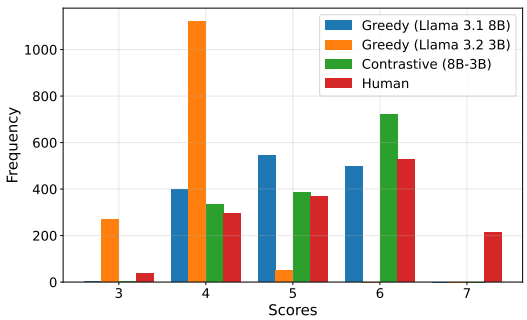
(a) Range 0-4



(b) Range 1-5

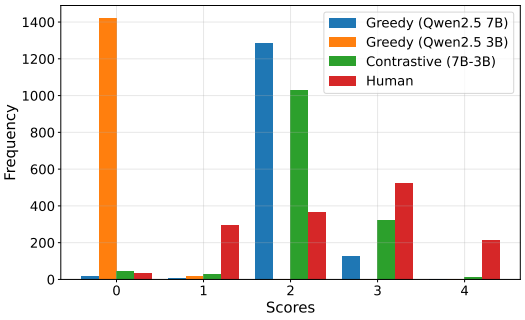


(c) Range 2-6

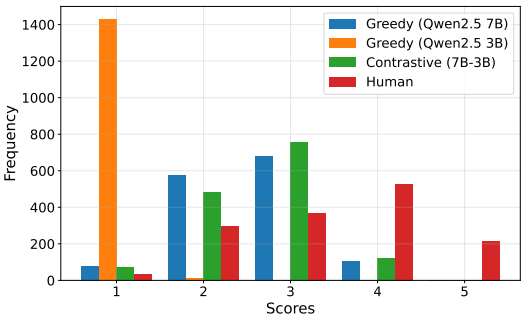


(d) Range 3-7

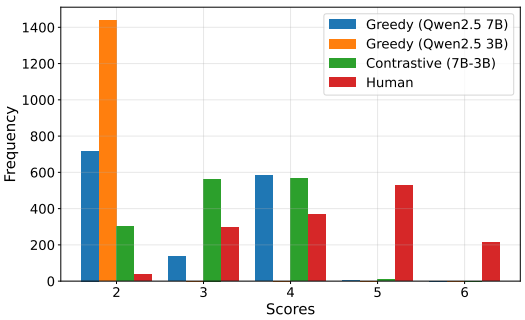
Figure 4: Predicted score distribution comparison for Llama-3 family across all score ranges on coherence evaluation.



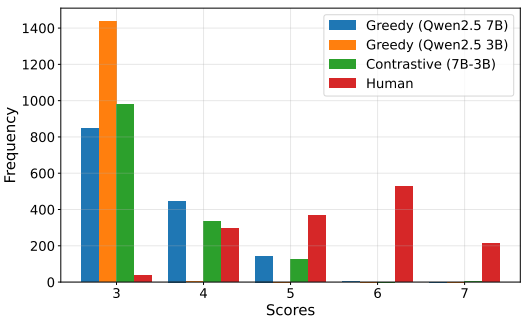
(a) Range 0-4



(b) Range 1-5

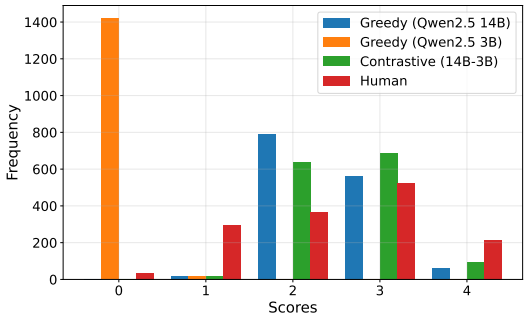


(c) Range 2-6

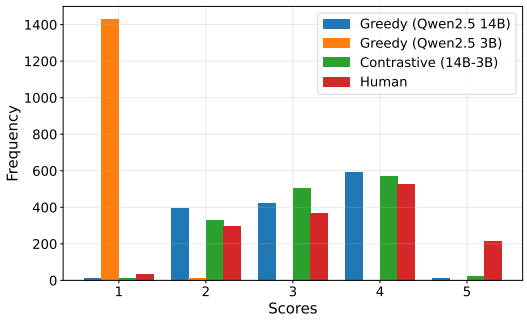


(d) Range 3-7

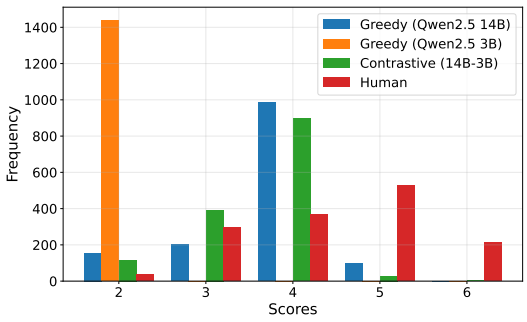
Figure 5: Predicted score distribution comparison for Qwen2.5 family across all score ranges on coherence evaluation.



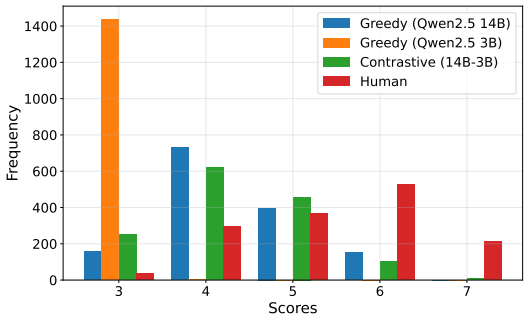
(a) Range 0-4



(b) Range 1-5



(c) Range 2-6



(d) Range 3-7

Figure 6: Score distribution comparison for Qwen2.5 14B family across all score ranges on coherence evaluation.

Model	Range	Pear.	Spear.	Kend.
Llama 3.2-1B	0 to 4	-.022$_{.025}$	.003$_{.038}$	.003$_{.025}$
Llama 3.2-3B	0 to 4	.051$_{.039}$	.073$_{.039}$	.068$_{.034}$
Llama 3.1-8B	0 to 4	.471$_{.044}$	.411$_{.027}$	.388$_{.030}$
Contra. (8B-3B)	0 to 4	.445$_{.039}$	.381$_{.030}$	.360$_{.027}$
Llama 3.2-1B	1 to 5	-.022$_{.007}$	-.028$_{.007}$	-.027$_{.009}$
Llama 3.2-3B	1 to 5	.178$_{.056}$	.196$_{.038}$	.182$_{.025}$
Llama 3.1-8B	1 to 5	.594$_{.047}$	.487$_{.053}$	.465$_{.032}$
Contra. (8B-3B)	1 to 5	.603$_{.056}$	.518$_{.048}$	.494$_{.035}$
Llama 3.2-1B	2 to 6	.000$_{.000}$	.000$_{.000}$	.000$_{.000}$
Llama 3.2-3B	2 to 6	.095$_{.054}$	.097$_{.059}$	.092$_{.037}$
Llama 3.1-8B	2 to 6	.488$_{.057}$	.431$_{.021}$	.413$_{.045}$
Contra. (8B-3B)	2 to 6	.551$_{.052}$	.469$_{.031}$	.447$_{.041}$
Llama 3.2-1B	3 to 7	-.011$_{.040}$	-.004$_{.050}$	-.004$_{.024}$
Llama 3.2-3B	3 to 7	.118$_{.035}$	.136$_{.051}$	.129$_{.038}$
Llama 3.1-8B	3 to 7	.549$_{.045}$	.484$_{.050}$	.452$_{.050}$
Contra. (8B-3B)	3 to 7	.540$_{.054}$	.496$_{.034}$	.465$_{.028}$
Average across all score ranges				
Llama 3.2-1B		-.014	-.007	-.007
Llama 3.2-3B		.111	.126	.118
Llama 3.1-8B		.526	.453	.430
Contra. (8B-3B)		**.535**	**.466**	**.442**

Table 12: Llama-3 family correlation results to human annotations on summary consistency. Max correlation within score range are underlined, max across all ranges are *italicized*, and max averages are **bolded**.

Model	Range	Pear.	Spear.	Kend.
Qwen2.5-3B	0 to 4	-.189$_{.085}$	-.192$_{.062}$	-.185$_{.051}$
Qwen2.5-7B	0 to 4	.396$_{.024}$	.392$_{.021}$	.356$_{.021}$
Qwen2.5-14B	0 to 4	.422$_{.023}$	.443$_{.036}$	.410$_{.034}$
Contra. (7B-3B)	0 to 4	.446$_{.025}$	.412$_{.028}$	.369$_{.035}$
Contra. (14B-3B)	0 to 4	.420$_{.038}$	.437$_{.041}$	.404$_{.035}$
Qwen2.5-3B	1 to 5	-.047$_{.071}$	-.063$_{.039}$	-.060$_{.055}$
Qwen2.5-7B	1 to 5	.542$_{.044}$	.479$_{.028}$	.444$_{.035}$
Qwen2.5-14B	1 to 5	.406$_{.051}$	.417$_{.049}$	.394$_{.037}$
Contra. (7B-3B)	1 to 5	.549$_{.044}$	.474$_{.033}$	.441$_{.030}$
Contra. (14B-3B)	1 to 5	.477$_{.051}$	.481$_{.045}$	.459$_{.031}$
Qwen2.5-3B	2 to 6	-.114$_{.045}$	-.095$_{.052}$	-.091$_{.062}$
Qwen2.5-7B	2 to 6	.489$_{.035}$	.435$_{.032}$	.404$_{.033}$
Qwen2.5-14B	2 to 6	.136$_{.034}$	.176$_{.028}$	.162$_{.024}$
Contra. (7B-3B)	2 to 6	.478$_{.021}$	.462$_{.038}$	.432$_{.029}$
Contra. (14B-3B)	2 to 6	.221$_{.029}$	.295$_{.034}$	.269$_{.034}$
Qwen2.5-3B	3 to 7	-.127$_{.074}$	-.126$_{.033}$	-.121$_{.049}$
Qwen2.5-7B	3 to 7	.391$_{.026}$	.414$_{.033}$	.367$_{.037}$
Qwen2.5-14B	3 to 7	.224$_{.046}$	.204$_{.027}$	.188$_{.030}$
Contra. (7B-3B)	3 to 7	.382$_{.028}$	.389$_{.036}$	.345$_{.036}$
Contra. (14B-3B)	3 to 7	.309$_{.033}$	.269$_{.043}$	.246$_{.034}$
Average across all score ranges				
Qwen2.5-3B		-.119	-.119	-.114
Qwen2.5-7B		.455	.430	.393
Qwen2.5-14B		.297	.310	.289
Contra. (7B-3B)		**.464**	**.434**	**.397**
Contra. (14B-3B)		.357	.370	.345

Table 13: Qwen2.5 family correlation results to human annotations on summary consistency. Max correlation within score range are underlined, max across all ranges are *italicized*, and max averages are **bolded**.

Range	Pearson	Spearman	Kendall
0-4	$.419_{.026}$	$.444_{.040}$	$.370_{.027}$
1-5	$.293_{.027}$	$.282_{.023}$	$.229_{.035}$
2-6	$.402_{.027}$	$.412_{.050}$	$.336_{.031}$
3-7	$.418_{.034}$	$.402_{.026}$	$.330_{.034}$
Average across all score ranges			
	.383	.385	.316

Table 14: Qwen2.5 14B-7B contrastive decoding results on summary coherence with 95% bootstrap confidence intervals. The 1-5 range shows significantly degraded performance compared to using a 3B assistant (Table 2), supporting the hypothesis that larger assistant models can penalize correct logits.