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

Anonymous ACL submission

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. 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, preventing the search for optimal 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 11.3% relative improvement on average in Spearman correlation with human judgments across different score ranges.

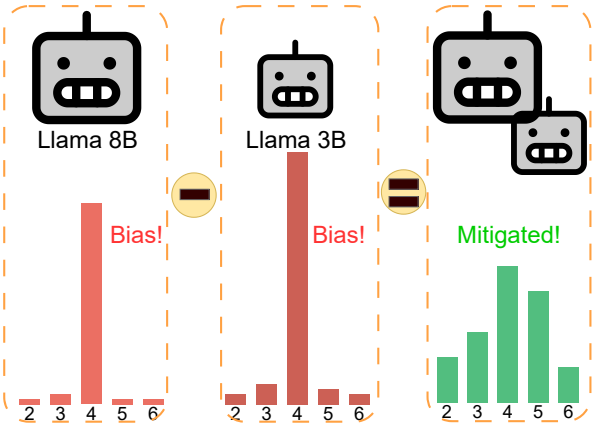


Figure 1: Overview of score range bias in 2-4 range and how contrastive decoding mitigates it through canceling out similar bias across models from the same family.

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?

In this work, we reveal another bias in LLM judge outputs, namely *score range biases*, where LLM judge outputs are sensitive to the shift in scores ranges, 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.

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.
- 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.

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* or *pointwise* 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 LLM judges assigning numerical ratings to outputs without any other output references. 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—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. Assuming these family biases are bidirectional, we aim to cancel out using contrastive decoding.

3 Contrastive Decoding for Mitigating LLM Judge Biases

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., $p_{\text{Asst}} = \frac{e_i/t}{\sum_j e_j/t}$. One difference we make compared to Li et al. (2023) is the inclusion of λ to further align the logit distribution between the two models motivated by our analysis in §4.2 .

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)[1]. 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.

Models We experiment on two model families.[2] 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, 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. We use 10% of the news articles are used as the held out development set.

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.

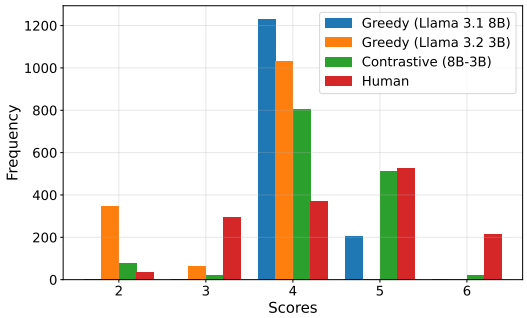
4.2 Reveal and Mitigate Score Range Biases

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

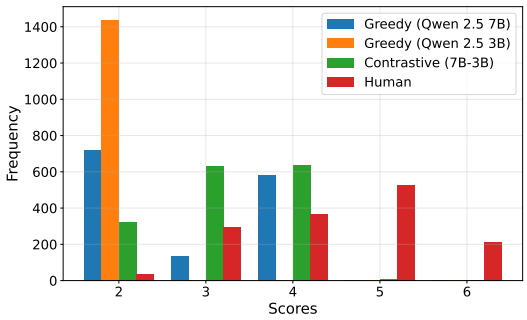
Upon analyzing the first output token logit distribution (Figure 3) of the Qwen family models in the 0-4 score range, Qwen-2.5 3B, 7B, 8B, and 14B

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

[2]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 learning and those models are finetuned towards 1-5 score range.

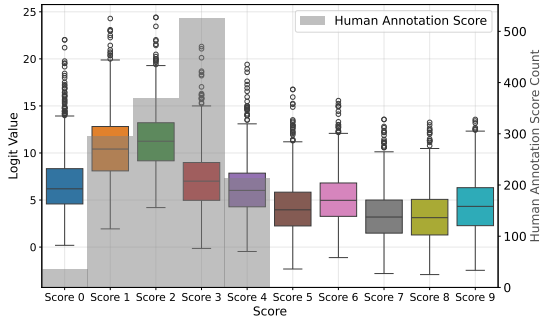


(a) Llama-3 Family Results

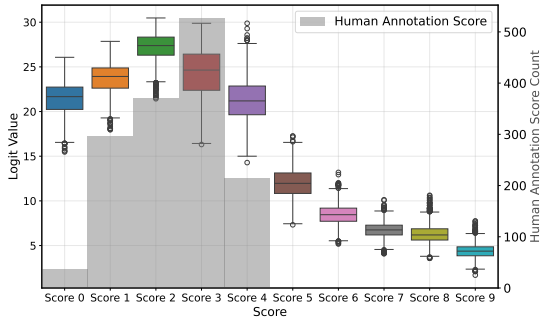


(b) Qwen2.5 Family Results

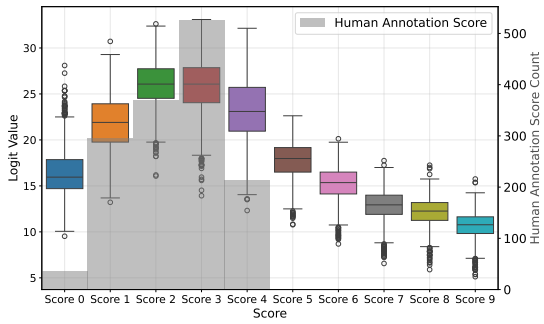
Figure 2: Coherence score distribution in 2-6 score range with greedy decoding, contrastive decoding, and human annotations. The greedy decoding outputs from both Llama 8B and 3B models are highly skewed towards outputting score of 4 and Qwen2.5 3B and 7B models are outputting score of 2 showing similar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



(a) Qwen2.5 3B logit distribution



(b) Qwen2.5 7B logit distribution



(c) Qwen2.5 14B logit distribution

Figure 3: Logit distribution of the first output token in 0-4 score range. For 3B and 7B models, the logit of Score 2 is the highest while the logit of Score 3 gets higher and becomes closer to human judgements as the model size increase from 3B through 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 ≈ 25 (Figure 3a), 7B ≈ 30 (Figure 3b), and 14B ≈ 34, further motivating the inclusion of λ in Eq. 1 to align the logit 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 score range from 1 to 5 shows the highest correlation among the experimented score ranges (7B, 1 to 5: 0.385, 14B, 1 to 5: 0.456), 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.[3]

Contrastive Decoding is a Robust Mitigation Strategy across Different Score Ranges Table 1 and 2 further show that contrastive decoding exhibits consistency in correlations across varying score ranges, addressing the score range bias observed in LLM judges. While using a single model suffers from decrease in correlation when score

[3]Also on 1-7 (Likert, 1932) or 1-10 (Ye et al., 2024) scales.

Model	Range	Pear.	Spear.	Kend.
Llama 3.2-1B-Inst	0 to 4	0.112	0.106	0.092
Llama 3.2-3B-Inst	0 to 4	0.141	0.146	0.126
Llama 3.1-8B-Inst	0 to 4	0.381	0.367	0.321
Contrastive (8B-1B)	0 to 4	0.386	0.369	0.324
Contrastive (8B-3B)	0 to 4	0.380	0.361	0.316
Llama 3.2-1B-Inst	1 to 5	0.071	0.069	0.060
Llama 3.2-3B-Inst	1 to 5	0.205	0.217	0.189
Llama 3.1-8B-Inst	1 to 5	0.347	0.338	0.293
Contrastive (8B-1B)	1 to 5	0.365	0.358	0.311
Contrastive (8B-3B)	1 to 5	0.340	0.336	0.293
Llama 3.2-1B-Inst	2 to 6	0.000	0.000	0.000
Llama 3.2-3B-Inst	2 to 6	0.168	0.167	0.148
Llama 3.1-8B-Inst	2 to 6	0.270	0.257	0.226
Contrastive (8B-1B)	2 to 6	0.310	0.302	0.264
Contrastive (8B-3B)	2 to 6	0.302	0.298	0.258
Llama 3.2-1B-Inst	3 to 7	-0.104	-0.128	-0.105
Llama 3.2-3B-Inst	3 to 7	0.033	0.045	0.040
Llama 3.1-8B-Inst	3 to 7	0.386	0.372	0.320
Contrastive (8B-1B)	3 to 7	0.383	*0.378*	*0.326*
Contrastive (8B-3B)	3 to 7	*0.393*	0.375	0.324
Average across all score ranges				
Llama 3.2-1B-Inst		0.020	0.012	0.012
Llama 3.2-3B-Inst		0.137	0.144	0.126
Llama 3.1-8B-Inst		0.346	0.334	0.290
Contrastive (8B-1B)		**0.361**	**0.352**	**0.306**
Contrastive (8B-3B)		0.354	0.343	0.298

Table 1: Llama-3 family correlation results to human annotations on summary coherence. Max correlation within score range are underlined, max across all ranges are *italicized*, and max averages are **bolded**. Maximal improvement is observed in the 2-6 score range.

Model	Range	Pear.	Spear.	Kend.
Qwen2.5-3B-Inst	0 to 4	-0.241	-0.221	-0.193
Qwen2.5-7B-Inst	0 to 4	0.264	0.266	0.234
Qwen2.5-14B-Inst	0 to 4	0.424	0.428	0.375
Contrastive (7B-3B)	0 to 4	0.330	0.333	0.291
Contrastive (14B-3B)	0 to 4	0.440	0.449	0.394
Qwen2.5-3B-Inst	1 to 5	-0.098	-0.086	-0.073
Qwen2.5-7B-Inst	1 to 5	0.385	0.385	0.333
Qwen2.5-14B-Inst	1 to 5	*0.460*	0.456	0.395
Contrastive (7B-3B)	1 to 5	0.360	0.351	0.304
Contrastive (14B-3B)	1 to 5	0.457	*0.459*	*0.397*
Qwen2.5-3B-Inst	2 to 6	-0.108	-0.086	-0.075
Qwen2.5-7B-Inst	2 to 6	0.373	0.363	0.312
Qwen2.5-14B-Inst	2 to 6	0.292	0.302	0.262
Contrastive (7B-3B)	2 to 6	0.377	0.363	0.312
Contrastive (14B-3B)	2 to 6	0.391	0.412	0.361
Qwen2.5-3B-Inst	3 to 7	0.301	0.313	0.273
Qwen2.5-7B-Inst	3 to 7	0.368	0.363	0.315
Qwen2.5-14B-Inst	3 to 7	0.354	0.350	0.304
Contrastive (7B-3B)	3 to 7	0.355	0.344	0.297
Contrastive (14B-3B)	3 to 7	0.407	0.411	0.353
Average across all score ranges				
Qwen2.5-3B-Inst		-0.036	-0.020	-0.017
Qwen2.5-7B-Inst		0.343	0.339	0.294
Qwen2.5-14B-Inst		0.383	0.384	0.334
Contrastive (7B-3B)		0.356	0.348	0.301
Contrastive (14B-3B)		**0.424**	**0.433**	**0.376**

Table 2: Qwen-2.5 family correlation results to human annotations on summary coherence. Max correlation within score range are underlined, max across all ranges are *italicized*, and max averages are **bolded**. Maximal improvement is observed in the 2-6 score range.

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 0.310 (compared to 0.168 for Llama 3.2-3B and 0.270 for Llama 3.1-8B) and similar improvements in Spearman and Kendall correlations, also seen as 5.1% relative improvement for Llama 8B ($0.335 \rightarrow 0.352$) and 11.3% for Qwen 14B ($0.389 \rightarrow 0.433$) on average across all score ranges. 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 C).

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 0.352 compared to 0.343 for the 3B assistant. However, the differences are small and depend on the evaluation dimension of summaries (Appendix C).

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. Both revealing and addressing these biases leads to better alignment of LLM judges with human evaluators and unlocking the potential to expand beyond the standard 1-5 score range. Future research directions include scaling to models beyond 14B parameters and tasks beyond summarization, and investigating alternative lightweight approaches e.g., prompt engineering on evaluation rubrics.

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 for speeding up decoding with speculative decoding (Leviathan et al., 2023), and therefore, contrastive decoding can be used without additional forward pass when speculative decoding is used.

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

Language Coverage Our experiments are conduced only on English language, however, we haven't exploited linguistic knowledge specific to English.

Task Coverage Our experiments are conducted on a summarization task following Liu et al. (2023) and Panickssery et al. (2024). However, we have conducted experiments on multiple dimensions of summarization metrics i.e., coherence, relevance (Appendix C), and consistency (Appendix C) to confirm score range bias is not happening with one specific dimension.

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, Sha-

ran 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, Xinfeng 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, 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*.

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 3: 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 4: Hyperparameter settings for contrastive decoding for each Llama-3 main and assistant model pair for relevance.

C Relevance and Consistency Results

In Tables 6 and 7, 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 8 and 9, we present the correlation results for summary consistency evaluation across different score ranges.

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

D 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[4] for Llama 3 family models and Apache-2.0 license for Qwen 2.5 family models. We followed their intended use case.

E 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.

F Potential Risks

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

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

Model	Range	Pear.	Spear.	Kend.
Llama 3.2-1B-Inst	0 to 4	0.336	0.317	0.283
Llama 3.2-3B-Inst	0 to 4	0.113	0.110	0.096
Llama 3.1-8B-Inst	0 to 4	0.420	0.385	*0.340*
Contrastive (8B-1B)	0 to 4	0.406	0.371	0.326
Contrastive (8B-3B)	0 to 4	0.403	0.370	0.327
Llama 3.2-1B-Inst	1 to 5	0.071	0.063	0.056
Llama 3.2-3B-Inst	1 to 5	0.138	0.148	0.130
Llama 3.1-8B-Inst	1 to 5	0.407	0.372	0.327
Contrastive (8B-1B)	1 to 5	0.399	0.374	0.331
Contrastive (8B-3B)	1 to 5	0.393	0.356	0.315
Llama 3.2-1B-Inst	2 to 6	0.000	0.000	0.000
Llama 3.2-3B-Inst	2 to 6	0.104	0.091	0.080
Llama 3.1-8B-Inst	2 to 6	0.305	0.288	0.255
Contrastive (8B-1B)	2 to 6	0.420	0.388	*0.340*
Contrastive (8B-3B)	2 to 6	0.325	0.305	0.270
Llama 3.2-1B-Inst	3 to 7	0.069	0.032	0.030
Llama 3.2-3B-Inst	3 to 7	0.099	0.096	0.082
Llama 3.1-8B-Inst	3 to 7	0.386	0.372	0.325
Contrastive (8B-1B)	3 to 7	0.395	0.376	0.330
Contrastive (8B-3B)	3 to 7	*0.429*	*0.396*	*0.340*
Average across all score ranges				
Llama 3.2-1B-Inst		0.119	0.103	0.092
Llama 3.2-3B-Inst		0.113	0.111	0.097
Llama 3.1-8B-Inst		0.379	0.354	0.312
Contrastive (8B-1B)		**0.405**	**0.377**	**0.332**
Contrastive (8B-3B)		0.388	0.357	0.313

Table 6: Llama-3 family correlation results to human annotations on summary relevance.

Model	Range	Pear.	Spear.	Kend.
Qwen2.5-3B-Inst	0 to 4	0.000	0.000	0.000
Qwen2.5-7B-Inst	0 to 4	0.304	0.282	0.247
Qwen2.5-14B-Inst	0 to 4	0.474	0.451	0.398
Contrastive (7B-3B)	0 to 4	0.329	0.284	0.246
Contrastive (14B-3B)	0 to 4	*0.502*	*0.487*	*0.425*
Qwen2.5-3B-Inst	1 to 5	0.000	0.000	0.000
Qwen2.5-7B-Inst	1 to 5	0.337	0.332	0.284
Qwen2.5-14B-Inst	1 to 5	0.489	0.467	0.411
Contrastive (7B-3B)	1 to 5	0.349	0.327	0.285
Contrastive (14B-3B)	1 to 5	0.469	0.448	0.394
Qwen2.5-3B-Inst	2 to 6	0.000	0.000	0.000
Qwen2.5-7B-Inst	2 to 6	0.232	0.222	0.195
Qwen2.5-14B-Inst	2 to 6	0.436	0.412	0.362
Contrastive (7B-3B)	2 to 6	0.247	0.261	0.225
Contrastive (14B-3B)	2 to 6	0.445	0.426	0.374
Qwen2.5-3B-Inst	3 to 7	0.000	0.000	0.000
Qwen2.5-7B-Inst	3 to 7	0.193	0.230	0.207
Qwen2.5-14B-Inst	3 to 7	0.448	0.419	0.366
Contrastive (7B-3B)	3 to 7	0.244	0.263	0.234
Contrastive (14B-3B)	3 to 7	0.483	0.469	0.407
Average across all score ranges				
Qwen2.5-3B-Inst		0.000	0.000	0.000
Qwen2.5-7B-Inst		0.267	0.267	0.233
Qwen2.5-14B-Inst		0.462	0.437	0.384
Contrastive (7B-3B)		0.292	0.284	0.248
Contrastive (14B-3B)		**0.475**	**0.458**	**0.400**

Table 7: Qwen2.5 family correlation results to human annotations on summary relevance.

Model	Range	Pear.	Spear.	Kend.
Llama 3.2-1B-Inst	0 to 4	-0.039	-0.027	-0.024
Llama 3.2-3B-Inst	0 to 4	0.118	0.115	0.110
Llama 3.1-8B-Inst	0 to 4	0.494	0.454	0.435
Contrastive (8B-1B)	0 to 4	0.482	0.435	0.418
Contrastive (8B-3B)	0 to 4	0.450	0.414	0.396
Llama 3.2-1B-Inst	1 to 5	-0.152	-0.229	-0.215
Llama 3.2-3B-Inst	1 to 5	0.208	0.213	0.202
Llama 3.1-8B-Inst	1 to 5	*0.675*	*0.581*	*0.560*
Contrastive (8B-1B)	1 to 5	0.672	0.578	0.557
Contrastive (8B-3B)	1 to 5	0.656	0.577	0.557
Llama 3.2-1B-Inst	2 to 6	0.000	0.000	0.000
Llama 3.2-3B-Inst	2 to 6	0.171	0.157	0.151
Llama 3.1-8B-Inst	2 to 6	0.476	0.449	0.434
Contrastive (8B-1B)	2 to 6	0.531	0.479	0.463
Contrastive (8B-3B)	2 to 6	0.548	0.497	0.480
Llama 3.2-1B-Inst	3 to 7	-0.000	0.007	0.008
Llama 3.2-3B-Inst	3 to 7	0.195	0.180	0.172
Llama 3.1-8B-Inst	3 to 7	0.555	0.509	0.483
Contrastive (8B-1B)	3 to 7	0.540	0.492	0.466
Contrastive (8B-3B)	3 to 7	0.550	0.510	0.488
Average across all score ranges				
Llama 3.2-1B-Inst		-0.048	-0.062	-0.058
Llama 3.2-3B-Inst		0.173	0.166	0.159
Llama 3.1-8B-Inst		0.550	0.498	0.478
Contrastive (8B-1B)		**0.557**	0.496	0.476
Contrastive (8B-3B)		0.551	**0.500**	**0.480**

Table 8: Llama-3 family correlation results to human annotations on summary consistency.

Model	Range	Pear.	Spear.	Kend.
Qwen2.5-3B-Inst	0 to 4	-0.486	-0.485	-0.474
Qwen2.5-7B-Inst	0 to 4	0.402	0.405	0.377
Qwen2.5-14B-Inst	0 to 4	0.452	0.456	0.430
Contrastive (7B-3B)	0 to 4	0.460	0.415	0.383
Contrastive (14B-3B)	0 to 4	0.451	0.453	0.428
Qwen2.5-3B-Inst	1 to 5	-0.282	-0.214	-0.208
Qwen2.5-7B-Inst	1 to 5	0.560	0.516	0.487
Qwen2.5-14B-Inst	1 to 5	0.482	0.479	0.459
Contrastive (7B-3B)	1 to 5	*0.565*	0.510	0.484
Contrastive (14B-3B)	1 to 5	0.564	*0.539*	*0.521*
Qwen2.5-3B-Inst	2 to 6	-0.175	-0.175	-0.170
Qwen2.5-7B-Inst	2 to 6	0.516	0.462	0.441
Qwen2.5-14B-Inst	2 to 6	0.165	0.171	0.158
Contrastive (7B-3B)	2 to 6	0.498	0.470	0.448
Contrastive (14B-3B)	2 to 6	0.280	0.313	0.292
Qwen2.5-3B-Inst	3 to 7	-0.456	-0.444	-0.432
Qwen2.5-7B-Inst	3 to 7	0.454	0.470	0.448
Qwen2.5-14B-Inst	3 to 7	0.260	0.244	0.233
Contrastive (7B-3B)	3 to 7	0.457	0.469	0.447
Contrastive (14B-3B)	3 to 7	0.343	0.306	0.289
Average across all score ranges				
Qwen2.5-3B-Inst		-0.350	-0.329	-0.321
Qwen2.5-7B-Inst		0.483	0.463	0.438
Qwen2.5-14B-Inst		0.340	0.338	0.320
Contrastive (7B-3B)		**0.495**	**0.466**	**0.441**
Contrastive (14B-3B)		0.408	0.403	0.382

Table 9: Qwen2.5 family correlation results to human annotations on summary consistency.