

Disentangling Severity and Centrality: Rater Effects in LLM-Based Automated Short-Answer Scoring

Xiaomeng Xiong Corinne Huggins-Manley Jinnie Shin

University of Florida

xiaomengxiong@ufl.edu amanley@coe.ufl.edu jinnie.shin@coe.ufl.edu

Abstract

In rater-effect research, centrality is often not modeled explicitly but instead inferred indirectly from score distributions or fit statistics. This study applies the Facets Model for Severity and Centrality (FM-SC) proposed by [Jin and Wang \(2018\)](#) to LLM-based automated scoring, jointly estimating severity and centrality as distinct rater effects. Using two Science and two ELA prompts from ASAP-SAS, we compared two human raters with 10 LLM raters spanning proprietary and open-weight models. Each LLM was evaluated under both zero-shot and few-shot prompting conditions.

Results showed that LLM centrality was strongly task-dependent rather than a stable model characteristic. The same LLM configurations showed little centrality on the Science prompts, whereas centrality was substantially more common on the ELA prompts. Few-shot prompting moved scale use closer to the human reference on some tasks, but this pattern was not consistent across prompts.

Comparisons between the standard MFRM and FM-SC showed that omitting differences in scale use altered severity estimates. The magnitude of scale-use departure was moderately positively correlated with the resulting severity shift ($r = .47$), with the largest shift approaching 1.5 logits. In addition, conventional MFRM fit diagnostics become more complex to interpret when evaluating rater effects in heterogeneous pools of LLM raters.

1 Introduction

Automated scoring has advanced rapidly in recent years, and some systems now show reasonably strong agreement with human raters. Quadratic weighted kappa (QWK; [Cohen, 1968](#)) is a common metric for evaluating such agreement and is often used as an operational criterion for acceptable scoring performance, such as the .70 threshold discussed by [Williamson et al. \(2012\)](#). On the ASAP-SAS dataset ([Barbara et al., 2012](#)), [Ormerod and](#)

[Kwako \(2024\)](#) found substantial variation in agreement with human scores when generative models were used without task-specific fine-tuning. Four small open-source generative models achieved average QWK values ranging from .23 to .43, whereas a previously reported GPT-4 benchmark reached .68. After task-specific fine-tuning on human-scored responses, the same small models achieved average QWK values of approximately .75. However, QWK indicates only the degree of agreement with human ratings, it does not reveal how models use the score scale.

Within a psychometric framework, researchers have used the many-facet Rasch model (MFRM) to examine such scale-use patterns by estimating systematic rater effects. [Jin et al. \(2025\)](#) found that LLMs were generally more severe than human raters and that few-shot prompting produced scores closer to human ratings than zero-shot prompting. [Yamashita \(2024\)](#) found that ChatGPT-4.0 had an estimated severity close to the average human rater but showed a compressed use of the 0–10 scale, assigning scores only between 3 and 7. [Jiao et al. \(2026\)](#) compared 10 LLMs with human experts and estimated model-specific severity parameters, showing that different LLMs exhibited distinct rater effects. [Wang et al. \(2025\)](#) combined generalizability theory and MFRM, extending the analysis beyond severity to include central tendency, range restriction, halo effects, and other rater effects.

Findings on central tendency have been mixed. Some studies reported score concentration for both human and LLM raters ([Wang et al., 2025](#)), whereas others found stronger avoidance of extreme categories among LLMs than among human raters (e.g., [Jin et al., 2025](#); [Yamashita, 2024](#); [Wu et al., 2026](#)). In contrast, [Jiao et al. \(2026\)](#) found slight evidence of central tendency among human raters but not among AI raters. Across these studies, however, central tendency was inferred indirectly from score distributions, scale use, or fit statistics

rather than modeled as a directly estimable rater effect. In a standard MFRM framework, where rater effects are represented only through severity, restricted scale use cannot be directly separated from rater severity.

Therefore, this study explicitly models central tendency in automated scoring as a separable rater parameter rather than a post hoc diagnostic pattern. We adopt the Facets Model for Severity and Centrality (FM-SC) proposed by Jin and Wang (2018), anchoring one human rater and estimating both severity and centrality for the second human rater and each LLM rater. Using four short-answer prompts from the ASAP-SAS dataset (Barbara et al., 2012), two Science and two ELA prompts with approximately 1,600–1,800 responses each, we compare 10 prompted generative LLMs in terms of their severity and central tendency relative to human raters. We further examine whether ignoring centrality affects severity estimates. The specific research questions are as follows:

RQ1. To what extent do prompted generative LLM raters exhibit centrality or extremity relative to human reference raters in short-answer scoring?

RQ2. How does LLM centrality vary across the four prompts, including the Science and ELA tasks?

RQ3. How does LLM centrality differ between zero-shot and few-shot prompting conditions across prompts?

RQ4. To what extent does omitting rater-specific scale use in a standard severity-only MFRM alter LLM rater severity estimates?

2 Methods

2.1 Data

The data were drawn from four short-answer items in the Automated Student Assessment Prize–Short Answer Scoring (ASAP-SAS) dataset (Barbara et al., 2012), covering two subject areas: two Science prompts (Prompts 1 and 10) and two ELA prompts (Prompts 8 and 9). The outcome was a holistic score. Prompt 1 used a four-point scale ranging from 0 to 3, whereas the other three prompts used a three-point scale ranging from 0 to 2.

For each prompt, the few-shot examples used for the LLMs were selected using deterministic rules: one response was selected from each score category, resulting in four examples for Prompt 1 and three examples for each of the other prompts. The

selected examples were required to show human-score consensus and typical response length. All scoring configurations used the same few-shot examples within each prompt. These examples were consistently excluded from the scoring analysis. The final analytic sample sizes were 1,668 for Prompt 1, 1,796 for Prompt 8, 1,795 for Prompt 9, and 1,637 for Prompt 10.

Each response received 22 ratings in a fully crossed design with no missing ratings: two human ratings (H1 and H2) and 20 LLM ratings from 10 generative models: GPT-4.1, Claude Sonnet 4.6, Gemma-4-31B, Mistral-Small-3.1, Llama-3.3-70B-Instruct, Nemotron-3-Nano-30B, Nemotron-3-Super-120B, GPT-OSS-20B, GPT-OSS-120B, and Granite-3.3-8B-Instruct. Each LLM was evaluated under zero-shot and few-shot prompting (Table 1). The two prompting configurations were treated as distinct raters to capture scoring differences induced by prompting strategy. All generative models were run under deterministic or near-deterministic settings. For models supporting explicit temperature control, temperature was set to 0. Internal reasoning was disabled for the Nemotron models, whereas the GPT-OSS models used the low reasoning-effort setting. Table 2 shows that the two human raters had highly similar score distributions and used all available categories across prompts.

2.2 Modeling Framework

The modeling framework is based on the many-facet Rasch model (MFRM; Linacre, 1989). In the standard adjacent-category formulation of the many-facet Rasch model, the log-odds of assigning ratee n on item i by rater r to category k rather than category $k - 1$ are given by

$$\log \frac{P(Y_{nir} = k)}{P(Y_{nir} = k - 1)} = \theta_n - \beta_i - \eta_r - \tau_k, \quad (1)$$

where θ_n denotes ratee ability, β_i denotes item difficulty, η_r denotes rater severity, and τ_k denotes the step threshold between categories $k - 1$ and k . In this formulation, rater effects are represented only through severity.

Jin and Wang (2018) showed that when raters differ in centrality, fitting the standard MFRM can produce biased estimates of rater severity and category thresholds, whereas explicitly modeling centrality substantially reduces this bias. Therefore, to model severity and centrality jointly, this study adopts the FM-SC model proposed by Jin and Wang (2018).

Because each prompt was modeled separately and contained a single holistic item, no separate item-difficulty facet was estimated. The FM-SC model adds a rater-specific centrality parameter ω_r to the threshold component:

$$\log \frac{P(Y_{nr} = k)}{P(Y_{nr} = k - 1)} = \theta_n - \eta_r - \omega_r \tau_k. \quad (2)$$

2.3 Identification and Baseline

For identification, the official resolved score (H1) was used as the reference rater, with $\eta_{H1} = 0$ and $\omega_{H1} = 1$ (Barbara et al., 2012). The second human rating (H2) was treated as a rater facet. Thus, rater severity was interpreted relative to H1: positive η_r values indicate greater severity than H1, whereas negative values indicate greater leniency than H1. Centrality was interpreted relative to H1: ω_r values above 1 indicate stronger centrality than H1, and values below 1 indicate greater extremity than H1. The fixed value $\omega_{H1} = 1$ serves only as the identification reference.

Following Jin and Wang (2018), individual ω_r estimates were not tested for significance. Guided by the .50 effect magnitudes used in the simulation study of Jin and Eckes (2022), we interpreted $|\log(\omega_r)| \geq .50$ as a notable scale-use departure ($\omega_r > 1$: centrality; $\omega_r < 1$: extremity) and $|\eta_r| \geq .50$ as notable severity ($\eta_r > 0$) or leniency ($\eta_r < 0$) relative to H1.

Because each prompt was calibrated separately, the absolute values of η and ω are not directly comparable across prompts. Rater effects were therefore interpreted within each prompt relative to the H1 reference.

2.4 Analysis

The analysis proceeded in three steps. First, FM-SC and the baseline MFRM were compared for each prompt using the deviance information criterion (DIC) to assess whether allowing rater-specific differences in threshold spacing improved model fit. Second, rater-level severity and centrality estimates from FM-SC were summarized across raters, prompts, and prompting conditions. Third, to examine the consequences of omitting rater-specific scale use, severity estimates from the two models were compared, with severity shifts calculated as the MFRM estimate minus the corresponding FM-SC estimate.

2.5 Estimation and Implementation

Models were estimated in a Bayesian framework using MCMC in JAGS through the `rjags` package. The estimation procedure followed Jin and Wang (2018), with adaptations for the prompt-specific design described above.

Weakly informative priors were assigned to all model parameters. Ratee ability was modeled as $\theta_n \sim N(0, \sigma_\theta^2)$ with $1/\sigma_\theta^2 \sim \text{Gamma}(0.1, 0.1)$; rater severity followed $\eta_r \sim N(0, 10)$; and centrality followed $\omega_r \sim \text{lognormal}(0, 1)$. For a prompt with C score categories, the first $C - 2$ step thresholds were freely estimated under the prior $\tau_k \sim N(0, 10)$. The final threshold was defined as the negative sum of the preceding thresholds, thereby constraining all step thresholds to sum to zero, $\sum_{k=1}^{C-1} \tau_k = 0$. This implementation follows the JAGS specification of Jin and Wang (2018), with the item-difficulty facet omitted because each prompt was modeled separately as a single holistic item.

For each prompt and model, three MCMC chains were run with 1,000 adaptation iterations, 5,000 burn-in iterations, and 5,000 retained iterations per chain. Posterior means were used as point estimates and posterior standard deviations as standard errors. Convergence diagnostics focused on the structural parameters η , ω , τ , and σ , with PSRF values below 1.10 treated as acceptable (Gelman et al., 2013). Model fit was compared using DIC. Analyses were conducted in R 4.3.2 using JAGS 4.3.2, `rjags` 4.17, and `coda` 0.19.4.1.

3 Results

3.1 Descriptive Patterns

Human agreement, defined by the H1–H2 QWK, was higher on the Science prompts than on the ELA prompts (Table 3). The mean QWK was .914 across P1 and P10 and .842 across P8 and P9. The generative LLMs showed the same overall pattern. Averaged across models and prompting conditions, mean QWK was .661 for Science and .508 for ELA. With the exception of Granite-3.3-8B, both Science prompts produced higher QWK values than both ELA prompts under both zero-shot and few-shot conditions.

Overall, Gemma-4-31B achieved the highest mean QWK (.682), followed closely by Claude Sonnet 4.6 (.677), Nemotron-Super-120B (.654), and GPT-4.1 (.653). This suggests that the strongest open-weight models reached levels of

agreement with human scores comparable to those of leading proprietary models. Nemotron-Nano-30B (.406) and Granite-3.3-8B (.401) showed the lowest overall agreement. The benefit of few-shot prompting was concentrated mainly on the ELA prompts. Mean QWK increased only slightly from .659 to .662 on Science, but increased from .482 to .534 on ELA, with the largest improvement on P9 from .497 to .575.

The score-category distributions showed a clearer concentration in the middle categories on the ELA prompts (Table 4). For P1, the two internal categories were combined for comparison. The mean proportion of LLM ratings assigned to the middle categories was .563 on P1 and .423 on P10, close to or below the corresponding H1 proportions of .570 and .470. In contrast, the mean middle-category proportions on P8 and P9 were .537 and .520, compared with .263 and .413 for H1. This pattern was not universal across models. On both P8 and P9, 4 of the 20 LLM configurations did not use the middle category most frequently and instead showed greater concentration at an endpoint, indicating substantial heterogeneity in scale use across models.

Table 5 reports infit and outfit mean-square statistics from the standard MFRM. Following Linacre's guidelines, values below 1.5 are considered productive for measurement, values between 1.5 and 2.0 are unproductive but not degrading, and values above 2.0 may distort measurement, as summarized by Myford and Wolfe (2003). Myford and Wolfe (2003) also noted that more stringent control limits of 0.8 to 1.2 may be appropriate in high-stakes settings. Overall, most generative LLM raters had mean-square values below 1, indicating less variation in their ratings than expected under the standard MFRM. A small number of rater-prompt combinations had values above 2.0, primarily involving Granite-3.3-8B, Nemotron-Nano-30B, and the two human raters on P8. Notably, H1 and H2 still showed high agreement on P8, with a QWK of .856, indicating that elevated fit statistics do not necessarily correspond to low interrater agreement. Because standard MFRM fit statistics indicate departures from model expectations without directly parameterizing differences in scale use, the following analyses use the FM-SC parameter ω to examine centrality and extremity relative to H1.

3.2 Model Comparison

Table 7 reports the DIC values for the FM-SC and standard MFRM models across the four prompts, and all key structural parameters showed satisfactory convergence (Table 6). The FM-SC model yielded lower DIC values than the MFRM for all four prompts, indicating improved model fit when rater-specific differences in threshold spacing were allowed. Better model fit alone, however, does not establish the presence of rater centrality. We therefore examine the direction and magnitude of scale-use differences below using the rater-specific ω parameters.

3.3 Rater Centrality and Severity Estimates

Centrality (ω). Figure 1 shows a strong prompt-dependent pattern in LLM scale use. H2 remained within the $|\log(\omega)| < .50$ range on all four prompts, whereas the 20 LLM configurations showed markedly different patterns across prompts. Using $|\log(\omega)| \geq .50$ as the interpretive benchmark, only one configuration showed notable centrality on each Science prompt: Nemotron-Nano-30B zero-shot on P1 ($\omega = 2.073$) and Mistral-Small-3.1 few-shot on P10 ($\omega = 1.741$). In contrast, centrality was much more common on the ELA prompts. On P8, 18 configurations showed notable centrality, with the largest values observed for Mistral-Small-3.1 zero-shot ($\omega = 4.435$), Mistral-Small-3.1 few-shot ($\omega = 4.334$), and Nemotron-Nano-30B zero-shot ($\omega = 4.131$); the two Granite-3.3-8B conditions instead showed notable extremity. On P9, 7 configurations showed notable centrality, with the largest values observed for GPT-4.1 zero-shot ($\omega = 2.363$), Mistral-Small-3.1 zero-shot ($\omega = 2.251$), and Mistral-Small-3.1 few-shot ($\omega = 2.086$). Overall, centrality was concentrated on the ELA prompts, particularly P8.

Centrality tended to be weaker under few-shot prompting, although this pattern was strongly prompt-dependent. P8 showed the most consistent pattern, with all 10 model families having ω estimates closer to H1 under few-shot than under zero-shot prompting. The same pattern was observed for 8 model families on both P1 and P9. In contrast, only 3 model families on P10 showed weaker centrality under few-shot prompting.

Severity (η). Severity estimates followed a different pattern from centrality (Figure 2). Whereas H2 remained close to the H1 reference across all four prompts, substantive severity departures were

common among the LLMs. Claude Sonnet 4.6 and GPT-4.1 were closest to H1 overall, with mean absolute severity deviations of 0.59 and 0.61 logits across prompts, respectively. Gemma-4-31B followed closely at 0.71 logits. In contrast, Granite-3.3-8B and Nemotron-Nano-30B each showed mean absolute deviations greater than 2.6 logits. Severity departures were slightly smaller overall under few-shot prompting, but this pattern was also prompt-dependent. Comparing absolute severity deviations, 7 models on P10 and 9 models on P9 were closer to H1 under few-shot than under zero-shot prompting, whereas P1 showed no consistent pattern and only 4 models on P8 moved closer to H1.

3.4 Effect of Omitting Centrality on Severity Estimates

Table 10 shows that severity estimates changed when rater-specific scale use was omitted from the model. Across the 80 LLM configuration-prompt combinations, $|\log(\omega)|$ was moderately positively correlated with $|\eta\text{-shift}|$ ($r = .47$), indicating that larger departures in scale use tended to be associated with larger changes in estimated severity. Configurations with $|\log(\omega)| \geq .50$ had a mean absolute severity shift of .376 logits, compared with .151 logits for configurations with smaller scale-use departures.

This pattern can also be seen within individual prompts. On P1, GPT-4.1 zero-shot was very close to the H1 scale-use reference ($\omega = 1.006$, $|\log(\omega)| = .006$) and had an η -shift of only .064, whereas Nemotron-Nano-30B zero-shot showed a larger scale-use departure ($\omega = 2.073$, $|\log(\omega)| = .729$) and a larger shift of .465. Similarly, on P8, Nemotron-Super-120B few-shot had $|\log(\omega)| = .762$ and an absolute severity shift of .067, whereas Mistral-Small-3.1 zero-shot had $|\log(\omega)| = 1.490$ and an absolute shift of 1.078.

4 Discussion

The most important finding of this study is that centrality was strongly task-dependent rather than a stable characteristic of an LLM rater. Prior studies have reported mixed evidence regarding central tendency and restricted scale use in LLM-based scoring (e.g., Jin et al., 2025; Yamashita, 2024; Wu et al., 2026; Jiao et al., 2026). The present results suggest that one reason for these inconsistent findings may be variation across scoring tasks. The

same 20 LLM configurations showed almost no notable centrality-side departures on the two Science tasks, whereas such departures were substantially more common on the ELA prompts. This suggests that centrality may be better understood as an interaction between the rater and the scoring task than as a general model characteristic. However, because the present study included only two prompts from each subject area, it cannot determine whether the observed differences reflect subject area, rubric structure, score-category clarity, or other prompt-specific characteristics.

Moreover, few-shot prompting may help calibrate how LLMs use the rating scale. Prior research suggests that the placement of scored anchors across the rating scale can affect calibration and midpoint clustering (Choi et al., 2026). One possible explanation is that scored examples operationalize otherwise abstract rubric categories by providing concrete score-point expectations, thereby helping models distinguish among score levels rather than defaulting to middle categories under uncertainty. In the present study, P8 showed the lowest mean LLM-human agreement across the four prompts, and few-shot prompting moved all 10 model families closer to the human reference in centrality. However, this pattern did not hold uniformly across the other prompts, indicating that the effect of few-shot prompting on scale use is itself task-dependent and may also depend on the selection of the examples provided.

The results also show that differences in LLM scale use cannot be adequately captured by severity alone. Across the 80 LLM configuration-prompt combinations, $|\log(\omega)|$ was moderately positively correlated with $|\eta\text{-shift}|$ ($r = .47$). Configurations reaching $|\log(\omega)| \geq .50$ had a mean absolute severity shift of .376, compared with .151 for the remaining configurations. Thus, stronger differences in scale use were associated with larger changes in severity estimates when centrality was omitted. Traditional MFRM primarily represents rater differences in terms of severity or leniency, whereas LLMs may also differ in how they use the available score categories. When such differences are not modeled separately, the severity parameter may reflect both rating location and scale use, potentially altering how an LLM is characterized as severe or lenient.

Also, this study highlights the need for caution when applying conventional rater-effect diagnostics to LLM raters, particularly with respect to the

context-dependent nature of fit statistics. Infit and outfit quantify the extent to which a rater's observed scoring pattern departs from the expectations of the fitted MFRM, and those expectations are defined by parameters jointly estimated from the full set of raters and responses. When the rater pool contains highly heterogeneous LLMs that differ substantially in scoring accuracy, scale use, and prompting conditions, the common measurement structure represented by the model may be strongly shaped by this heterogeneity. In the present study, for example, most of the 20 LLM configurations on P8 showed substantial concentration in the middle score category, while both human raters had infit and outfit values above 2 despite maintaining relatively high agreement with one another ($QWK = .856$). This pattern suggests that, in a heterogeneous human–LLM rater pool, fit statistics should not be interpreted as direct indicators of rater quality or of a specific rater effect such as centrality. The interpretation of conventional MFRM fit diagnostics may become more complex when they are extended to LLM-as-rater applications.

References

- Barbara, Ben Hamner, Jaison Morgan, lynnvande, and Mark Shermis. 2012. The hewlett foundation: Short answer scoring. <https://kaggle.com/competitions/asap-sas>. Kaggle.
- Jaeyoon Choi, Tamara Tate, and Mark Warschauer. 2026. Anchor is the key: Toward accessible automated essay scoring with large language model through prompting. *Assessing Writing*, 69:101053.
- Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213.
- Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. 2013. *Bayesian Data Analysis*. CRC Press.
- Hong Jiao, Dan Song, and Won-Chan Lee. 2026. Evaluating rater effects of large language models in automated essay scoring: Gpt, claude, gemini, and deepseek. *Educational Measurement: Issues and Practice*, 45(2):e70018.
- Kuan-Yu Jin and Thomas Eckes. 2022. Detecting differential rater functioning in severity and centrality: The dual drf facets model. *Educational and Psychological Measurement*, 82(4):757–781.
- Kuan-Yu Jin and Wen-Chung Wang. 2018. A new facets model for rater’s centrality/extremity response style. *Journal of Educational Measurement*, 55(4):543–563.
- Rui Jin, Mingren Zhao, Chunling Niu, Yuyan Xia, Hao Zhou, and Na Liu. 2025. Evaluating the performance of chatgpt and claude in automated writing scoring: Insights from the many-facet rasch model. *Education and Information Technologies*, 30(18):25881–25908.
- John Michael Linacre. 1989. *Many-faceted Rasch measurement*. Ph.D. thesis, The University of Chicago.
- Carol M Myford and Edward W Wolfe. 2003. Detecting and measuring rater effects using many-facet rasch measurement: Part i. *Journal of applied measurement*, 4(4):386–422.
- Christopher Michael Ormerod and Alexander Kwako. 2024. Automated text scoring in the age of generative ai for the gpu-poor. *arXiv preprint arXiv:2407.01873*.
- Yuehan Wang, Jinyan Huang, Lun Du, Yuxin Guo, Ying Liu, and Rong Wang. 2025. Evaluating large language models as raters in large-scale writing assessments: A psychometric framework for reliability and validity. *Computers and Education: Artificial Intelligence*, page 100481.
- David M Williamson, Xiaoming Xi, and F Jay Breyer. 2012. A framework for evaluation and use of automated scoring. *Educational measurement: issues and practice*, 31(1):2–13.
- Hsiang-Ning Wu, Man-Ni Chu, and Jia-Lien Hsu. 2026. Comparing gpt and human raters in essay assessment: Variability, bias, and the potential of llm-based scoring. *Computers and Education Open*, page 100341.
- Taichi Yamashita. 2024. An application of many-facet rasch measurement to evaluate automated essay scoring: A case of chatgpt-4.0. *Research Methods in Applied Linguistics*, 3(3):100133.

A Appendix

Table 1: Raters and generative models used in the study

Rater / Model	Parameters	Developer	Rater conditions
<i>Human raters</i>			
H1	–	–	Human rating
H2	–	–	Human rating
<i>Proprietary generative LLMs</i>			
GPT-4.1	–	OpenAI	Zero-shot; few-shot
Claude Sonnet 4.6	–	Anthropic	Zero-shot; few-shot
<i>Open-weight generative LLMs</i>			
Nemotron-3-Super-120B-A12B	120B / 12B active	NVIDIA	Zero-shot; few-shot
GPT-OSS-120B	117B / 5.1B active	OpenAI	Zero-shot; few-shot
Llama-3.3-70B-Instruct	70B	Meta	Zero-shot; few-shot
Nemotron-3-Nano-30B-A3B	31.6B / 3.2B active	NVIDIA	Zero-shot; few-shot
Gemma-4-31B-IT	31B	Google	Zero-shot; few-shot
Mistral-Small-3.1	24B	Mistral AI	Zero-shot; few-shot
GPT-OSS-20B	21B / 3.6B active	OpenAI	Zero-shot; few-shot
Granite-3.3-8B-Instruct	8B	IBM	Zero-shot; few-shot

Note. Each generative model contributes two distinct rater conditions, one zero-shot and one few-shot, yielding 20 LLM raters and 22 raters in total including H1 and H2. “–” indicates that parameter counts are not publicly disclosed or not applicable.

Table 2: Descriptive statistics of human ratings (H1, H2) by prompt

Prompt	Rater	N	Mean	SD	Score category (%)			
					0	1	2	3
Science P1	H1	1668	1.49	1.05	22.7	25.7	31.4	20.3
	H2	1668	1.48	1.06	23.0	25.7	31.2	20.1
Science P10	H1	1637	1.18	0.71	17.7	47.0	35.4	–
	H2	1637	1.16	0.70	17.9	48.3	33.8	–
ELA P8	H1	1796	1.13	0.85	30.5	26.3	43.2	–
	H2	1796	1.13	0.85	30.7	25.9	43.4	–
ELA P9	H1	1795	1.10	0.76	24.1	41.3	34.6	–
	H2	1795	1.10	0.76	24.7	40.7	34.6	–

Note. Score categories range from 0–3 for Science P1 and 0–2 for the other prompts.

Table 3: Quadratic weighted kappa (QWK) between each rater by prompt

Rater	Science		ELA	
	P1	P10	P8	P9
H1 vs H2 (human reference)	0.944	0.884	0.856	0.828
Claude Sonnet 4.6 few-shot	0.802	0.734	0.565	0.664
Claude Sonnet 4.6 zero-shot	0.791	0.727	0.542	0.589
GPT-4.1 few-shot	0.778	0.757	0.588	0.648
GPT-4.1 zero-shot	0.785	0.748	0.476	0.441
Gemma-4-31B few-shot	0.788	0.740	0.586	0.636
Gemma-4-31B zero-shot	0.799	0.754	0.589	0.562
Granite-3.3-8B few-shot	0.537	0.264	0.345	0.400
Granite-3.3-8B zero-shot	0.608	0.252	0.407	0.395
Llama-3.3-70B few-shot	0.740	0.644	0.512	0.625
Llama-3.3-70B zero-shot	0.730	0.573	0.480	0.556
Mistral-Small-3.1 few-shot	0.730	0.574	0.390	0.517
Mistral-Small-3.1 zero-shot	0.734	0.529	0.367	0.432
Nemotron-Nano-30B few-shot	0.422	0.490	0.322	0.409
Nemotron-Nano-30B zero-shot	0.495	0.474	0.344	0.294
Nemotron-Super-120B few-shot	0.698	0.715	0.606	0.643
Nemotron-Super-120B zero-shot	0.696	0.699	0.549	0.625
GPT-OSS-20B few-shot	0.690	0.694	0.495	0.599
GPT-OSS-20B zero-shot	0.681	0.695	0.431	0.550
GPT-OSS-120B few-shot	0.772	0.678	0.522	0.605
GPT-OSS-120B zero-shot	0.742	0.674	0.481	0.528

Note. QWK is computed between each rater and H1.

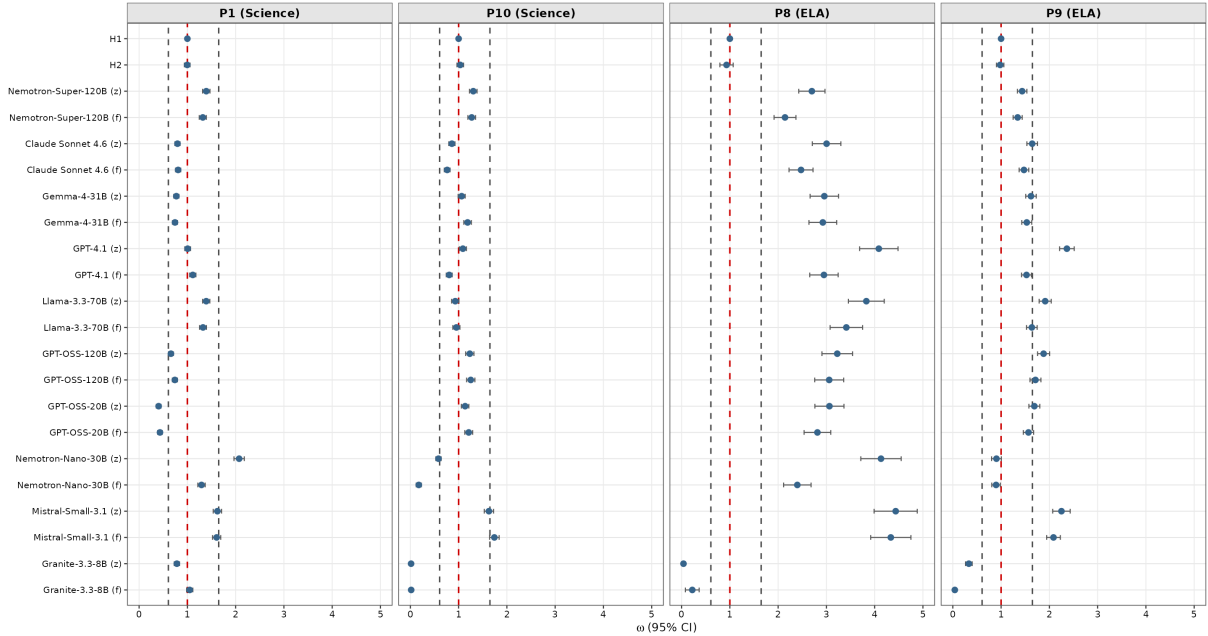


Figure 1: Rater-specific scale-use estimates (ω) by prompt. Estimates above 1 indicate centrality relative to H1; values below 1 indicate extremity. The red dashed line marks the H1 reference at $\omega = 1$. The gray dashed lines mark the literature-informed benchmark corresponding to $|\log(\omega)| = 0.5$, or approximately $\omega = 0.61$ and $\omega = 1.65$.

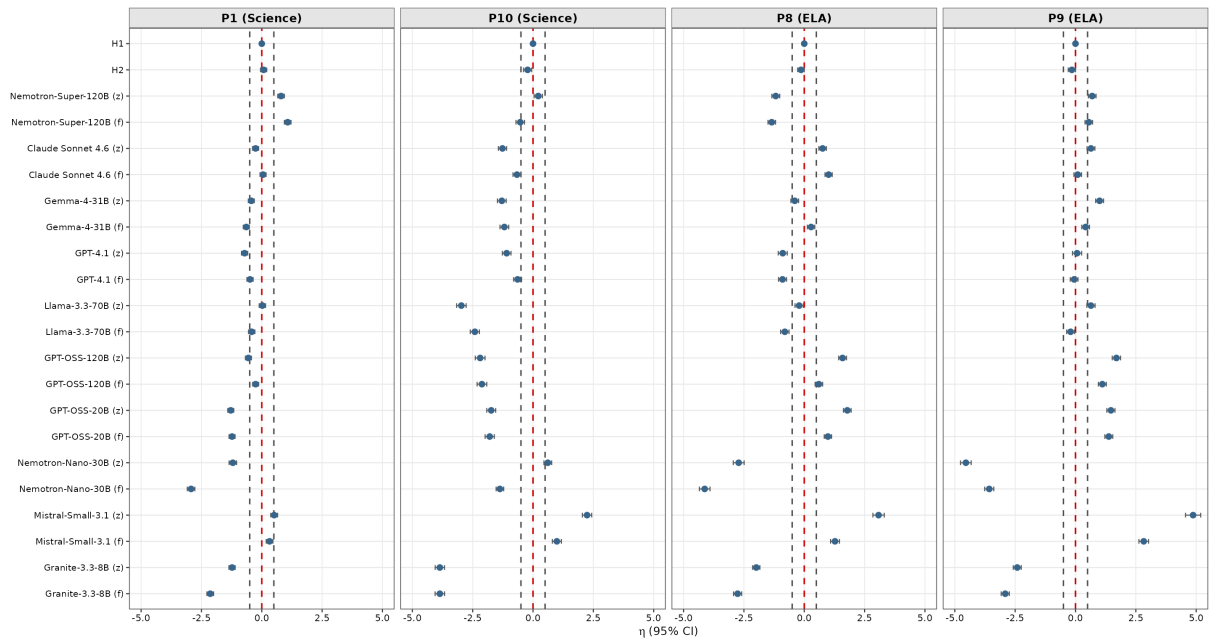


Figure 2: Rater severity estimates (η) by prompt. The red dashed line marks the H1 reference at $\eta = 0$. Positive values indicate greater severity relative to H1; negative values indicate greater leniency. The gray dashed lines mark the half-logit benchmark at $\eta = -0.5$ and $\eta = 0.5$.

Table 4: Full score category distributions

Rater	P1 (Science)				P10 (Science)			P8 (ELA)			P9 (ELA)		
	0	1	2	3	0	1	2	0	1	2	0	1	2
H1	0.227	0.257	0.313	0.203	0.177	0.470	0.354	0.305	0.263	0.432	0.241	0.413	0.346
H2	0.230	0.257	0.312	0.202	0.179	0.482	0.338	0.307	0.259	0.434	0.247	0.407	0.346
Claude Sonnet 4.6 few-shot	0.253	0.257	0.236	0.255	0.199	0.384	0.418	0.280	0.523	0.198	0.191	0.563	0.246
Claude Sonnet 4.6 zero-shot	0.236	0.235	0.247	0.282	0.141	0.415	0.445	0.216	0.608	0.176	0.202	0.611	0.187
GPT-4.1 few-shot	0.146	0.327	0.287	0.240	0.192	0.400	0.408	0.125	0.562	0.312	0.174	0.577	0.248
GPT-4.1 zero-shot	0.151	0.299	0.266	0.283	0.118	0.495	0.388	0.080	0.710	0.210	0.090	0.778	0.132
Gemma-4-31B few-shot	0.196	0.272	0.187	0.345	0.101	0.523	0.376	0.189	0.595	0.216	0.204	0.580	0.216
Gemma-4-31B zero-shot	0.205	0.287	0.188	0.321	0.111	0.482	0.407	0.149	0.584	0.267	0.230	0.600	0.170
Granite-3.3-8B few-shot	0.134	0.082	0.423	0.361	0.113	0.016	0.870	0.146	0.094	0.761	0.170	0.070	0.760
Granite-3.3-8B zero-shot	0.225	0.055	0.383	0.337	0.118	0.007	0.876	0.217	0.059	0.724	0.170	0.155	0.675
Llama-3.3-70B few-shot	0.153	0.261	0.423	0.162	0.076	0.404	0.520	0.109	0.631	0.261	0.152	0.606	0.242
Llama-3.3-70B zero-shot	0.169	0.288	0.423	0.120	0.059	0.363	0.578	0.115	0.704	0.182	0.165	0.681	0.154
Mistral-Small-3.1 few-shot	0.185	0.240	0.526	0.049	0.134	0.713	0.153	0.158	0.769	0.072	0.291	0.656	0.053
Mistral-Small-3.1 zero-shot	0.187	0.267	0.502	0.044	0.246	0.643	0.112	0.275	0.702	0.023	0.443	0.542	0.014
Nemotron-Nano-30B few-shot	0.071	0.113	0.436	0.381	0.241	0.169	0.591	0.043	0.256	0.700	0.074	0.245	0.681
Nemotron-Nano-30B zero-shot	0.078	0.196	0.669	0.057	0.355	0.287	0.358	0.036	0.602	0.362	0.046	0.196	0.758
Nemotron-Super-120B few-shot	0.206	0.434	0.251	0.109	0.122	0.562	0.316	0.143	0.416	0.441	0.242	0.523	0.235
Nemotron-Super-120B zero-shot	0.181	0.414	0.298	0.106	0.160	0.577	0.263	0.125	0.511	0.364	0.237	0.550	0.212
GPT-OSS-20B few-shot	0.225	0.153	0.154	0.468	0.074	0.511	0.415	0.247	0.578	0.175	0.268	0.577	0.155
GPT-OSS-20B zero-shot	0.228	0.141	0.154	0.478	0.084	0.491	0.425	0.292	0.596	0.112	0.254	0.609	0.137
GPT-OSS-120B few-shot	0.243	0.235	0.225	0.297	0.060	0.510	0.430	0.201	0.616	0.183	0.224	0.623	0.153
GPT-OSS-120B zero-shot	0.240	0.200	0.218	0.342	0.059	0.501	0.440	0.261	0.626	0.113	0.239	0.654	0.107

Note. Each cell reports the proportion of ratings assigned to that score category.

Table 5: Infit and outfit mean-square statistics for each rater from the standard MFRM, by prompt

Prompt	Rater	Infit MSQ	Outfit MSQ
Science P1	H1	1.043	0.956
	H2	1.031	0.942
	Claude Sonnet 4.6 zero-shot	0.595	0.532
	Claude Sonnet 4.6 few-shot	0.583	0.520
	GPT-4.1 zero-shot	0.502	0.498
	GPT-4.1 few-shot	0.532	0.571
	Gemma-4-31B zero-shot	0.665	0.593
	Gemma-4-31B few-shot	0.683	0.600
	Granite-3.3-8B zero-shot	1.884	1.651
	Granite-3.3-8B few-shot	1.807	1.773
	Llama-3.3-70B zero-shot	0.577	0.580
	Llama-3.3-70B few-shot	0.553	0.589
	Mistral-Small-3.1 zero-shot	0.663	0.708
	Mistral-Small-3.1 few-shot	0.606	0.618
	Nemotron-Nano-30B zero-shot	1.599	2.136
	Nemotron-Nano-30B few-shot	1.951	2.325
	Nemotron-Super-120B zero-shot	0.631	0.654
	Nemotron-Super-120B few-shot	0.655	0.709
	GPT-OSS-20B zero-shot	1.307	1.044
	GPT-OSS-20B few-shot	1.310	1.054
	GPT-OSS-120B zero-shot	1.005	0.899
	GPT-OSS-120B few-shot	0.921	0.830
Science P10	H1	0.905	0.750
	H2	1.003	0.899
	Claude Sonnet 4.6 zero-shot	0.683	0.518
	Claude Sonnet 4.6 few-shot	0.712	0.538
	GPT-4.1 zero-shot	0.500	0.374
	GPT-4.1 few-shot	0.692	0.532
	Gemma-4-31B zero-shot	0.425	0.301
	Gemma-4-31B few-shot	0.423	0.314
	Granite-3.3-8B zero-shot	2.917	1.441
	Granite-3.3-8B few-shot	2.918	1.723
	Llama-3.3-70B zero-shot	0.686	0.497
	Llama-3.3-70B few-shot	0.630	0.427
	Mistral-Small-3.1 zero-shot	0.879	0.780
	Mistral-Small-3.1 few-shot	0.800	0.706
	Nemotron-Nano-30B zero-shot	2.128	1.942
	Nemotron-Nano-30B few-shot	2.737	2.178
	Nemotron-Super-120B zero-shot	0.586	0.457
	Nemotron-Super-120B few-shot	0.574	0.419
	GPT-OSS-20B zero-shot	0.526	0.382
	GPT-OSS-20B few-shot	0.487	0.358
	GPT-OSS-120B zero-shot	0.493	0.357
	GPT-OSS-120B few-shot	0.466	0.341
ELA P8	H1	2.121	2.143
	H2	2.073	2.069
	Claude Sonnet 4.6 zero-shot	0.591	0.481
	Claude Sonnet 4.6 few-shot	0.675	0.547
	GPT-4.1 zero-shot	0.611	0.623
	GPT-4.1 few-shot	0.498	0.386
	Gemma-4-31B zero-shot	0.559	0.466
	Gemma-4-31B few-shot	0.514	0.398
	Granite-3.3-8B zero-shot	2.047	1.649
	Granite-3.3-8B few-shot	2.017	1.451
	Llama-3.3-70B zero-shot	0.581	0.491
	Llama-3.3-70B few-shot	0.556	0.462
	Mistral-Small-3.1 zero-shot	0.632	0.531
	Mistral-Small-3.1 few-shot	0.662	0.610
	Nemotron-Nano-30B zero-shot	1.152	1.304
	Nemotron-Nano-30B few-shot	1.074	0.911
	Nemotron-Super-120B zero-shot	0.639	0.530
	Nemotron-Super-120B few-shot	0.664	0.535
	GPT-OSS-20B zero-shot	0.954	0.895
	GPT-OSS-20B few-shot	0.934	0.841
	GPT-OSS-120B zero-shot	0.785	0.696
	GPT-OSS-120B few-shot	0.667	0.554
ELA P9	H1	1.629	1.814
	H2	1.603	1.680
	Claude Sonnet 4.6 zero-shot	0.659	0.521
	Claude Sonnet 4.6 few-shot	0.540	0.390
	GPT-4.1 zero-shot	0.896	1.506
	GPT-4.1 few-shot	0.544	0.398
	Gemma-4-31B zero-shot	0.751	0.627
	Gemma-4-31B few-shot	0.598	0.444
	Granite-3.3-8B zero-shot	2.044	1.834
	Granite-3.3-8B few-shot	2.299	1.756
	Llama-3.3-70B zero-shot	0.597	0.502
	Llama-3.3-70B few-shot	0.580	0.424
	Mistral-Small-3.1 zero-shot	0.603	0.496
	Mistral-Small-3.1 few-shot	0.605	0.489
	Nemotron-Nano-30B zero-shot	1.519	1.083
	Nemotron-Nano-30B few-shot	1.227	1.000
	Nemotron-Super-120B zero-shot	0.698	0.525
	Nemotron-Super-120B few-shot	0.733	0.574
	GPT-OSS-20B zero-shot	0.871	0.839
	GPT-OSS-20B few-shot	0.773	0.671
	GPT-OSS-120B zero-shot	0.816	0.833
	GPT-OSS-120B few-shot	0.655	0.511

Table 6: MCMC convergence diagnostics

Prompt	Maximum PSRF	
	FM-SC	MFRM
P1 (Science)	1.005	1.002
P10 (Science)	1.022	1.015
P8 (ELA)	1.040	1.013
P9 (ELA)	1.010	1.002

Note. Entries are the maximum PSRF across the structural parameters for each prompt. For FM-SC, these include η_r , ω_r , τ_k , and σ . For MFRM, they include η_r , τ_k , and σ .

Table 7: Model comparison based on the deviance information criterion (DIC) between FM-SC and MFRM, by prompt

Prompt	DIC		Preferred
	FM-SC	MFRM	
P1 (Science)	48,563	54,127	FM-SC
P10 (Science)	29,883	35,420	FM-SC
P8 (ELA)	34,841	41,993	FM-SC
P9 (ELA)	32,936	37,705	FM-SC

Note. Lower DIC values indicate better model fit.

Table 8: Estimated step parameters (τ_k) by prompt and model

Prompt	Model	τ_1	τ_2	τ_3
P1	MFRM	-2.793 (.031)	-0.083 (.019)	2.876 (.030)
	FM-SC	-2.824 (.064)	-0.294 (.020)	3.118 (.070)
P10	MFRM	-2.399 (.024)	2.399 (.024)	—
	FM-SC	-2.753 (.069)	2.753 (.069)	—
P8	MFRM	-2.600 (.024)	2.600 (.024)	—
	FM-SC	-1.096 (.052)	1.096 (.052)	—
P9	MFRM	-2.875 (.025)	2.875 (.025)	—
	FM-SC	-2.246 (.069)	2.246 (.069)	—

Note. Values are posterior means with posterior standard deviations in parentheses. Step parameters were estimated under the identification constraint $\sum_k \tau_k = 0$. P1 has four score categories and therefore three step parameters; P8, P9, and P10 have three score categories and therefore two step parameters.

Table 9: Raw rating scores and rater parameter estimates (η , ω) by prompt

Prompt	Rater	Raw rating		Severity	Centrality
		Mean	SD	η (SD)	ω (SD)
Science P1	H1	1.49	1.05	0.00*	1.00*
	H2	1.48	1.06	0.076 (0.065)	0.997 (0.030)
	Claude Sonnet 4.6 zero-shot	1.58	1.13	-0.257 (0.063)	0.795 (0.027)
	Claude Sonnet 4.6 few-shot	1.49	1.13	0.053 (0.063)	0.808 (0.028)
	GPT-4.1 zero-shot	1.68	1.04	-0.712 (0.066)	1.006 (0.030)
	GPT-4.1 few-shot	1.62	1.00	-0.487 (0.065)	1.112 (0.033)
	Gemma-4-31B zero-shot	1.62	1.13	-0.434 (0.063)	0.772 (0.027)
	Gemma-4-31B few-shot	1.68	1.14	-0.641 (0.064)	0.744 (0.026)
	Granite-3.3-8B zero-shot	1.83	1.12	-1.231 (0.066)	0.782 (0.028)
	Granite-3.3-8B few-shot	2.01	0.99	-2.135 (0.070)	1.048 (0.033)
	Llama-3.3-70B zero-shot	1.50	0.91	0.025 (0.067)	1.391 (0.038)
	Llama-3.3-70B few-shot	1.59	0.93	-0.418 (0.067)	1.324 (0.037)
	Mistral-Small-3.1 zero-shot	1.40	0.84	0.514 (0.072)	1.622 (0.044)
	Mistral-Small-3.1 few-shot	1.44	0.85	0.322 (0.071)	1.606 (0.043)
	Nemotron-Nano-30B zero-shot	1.70	0.69	-1.199 (0.079)	2.073 (0.055)
	Nemotron-Nano-30B few-shot	2.13	0.87	-2.930 (0.080)	1.293 (0.039)
	Nemotron-Super-120B zero-shot	1.33	0.89	0.798 (0.069)	1.392 (0.039)
	Nemotron-Super-120B few-shot	1.26	0.91	1.073 (0.070)	1.320 (0.037)
	GPT-OSS-20B zero-shot	1.88	1.23	-1.285 (0.064)	0.404 (0.025)
	GPT-OSS-20B few-shot	1.86	1.23	-1.231 (0.063)	0.433 (0.025)
	GPT-OSS-120B zero-shot	1.66	1.18	-0.553 (0.064)	0.659 (0.026)
	GPT-OSS-120B few-shot	1.58	1.15	-0.254 (0.063)	0.741 (0.027)
Science P10	H1	1.18	0.71	0.00*	1.00*
	H2	1.16	0.70	-0.222 (0.085)	1.034 (0.035)
	Claude Sonnet 4.6 zero-shot	1.30	0.70	-1.265 (0.086)	0.861 (0.033)
	Claude Sonnet 4.6 few-shot	1.22	0.75	-0.664 (0.084)	0.763 (0.030)
	GPT-4.1 zero-shot	1.27	0.66	-1.092 (0.091)	1.088 (0.037)
	GPT-4.1 few-shot	1.22	0.74	-0.638 (0.084)	0.806 (0.031)
	Gemma-4-31B zero-shot	1.30	0.66	-1.292 (0.092)	1.063 (0.037)
	Gemma-4-31B few-shot	1.28	0.63	-1.187 (0.092)	1.186 (0.040)
	Granite-3.3-8B zero-shot	1.76	0.65	-3.862 (0.098)	0.015 (0.007)
	Granite-3.3-8B few-shot	1.76	0.64	-3.861 (0.100)	0.017 (0.007)
	Llama-3.3-70B zero-shot	1.52	0.61	-2.969 (0.101)	0.930 (0.038)
	Llama-3.3-70B few-shot	1.44	0.63	-2.410 (0.096)	0.954 (0.037)
	Mistral-Small-3.1 zero-shot	0.87	0.58	2.239 (0.098)	1.628 (0.049)
	Mistral-Small-3.1 few-shot	1.02	0.54	0.989 (0.096)	1.741 (0.050)
	Nemotron-Nano-30B zero-shot	1.00	0.84	0.613 (0.082)	0.583 (0.028)
	Nemotron-Nano-30B few-shot	1.35	0.84	-1.372 (0.080)	0.175 (0.027)
	Nemotron-Super-120B zero-shot	1.10	0.64	0.216 (0.087)	1.304 (0.040)
	Nemotron-Super-120B few-shot	1.19	0.63	-0.529 (0.090)	1.271 (0.040)
	GPT-OSS-20B zero-shot	1.34	0.63	-1.732 (0.094)	1.135 (0.039)
	GPT-OSS-20B few-shot	1.34	0.61	-1.795 (0.097)	1.210 (0.041)
	GPT-OSS-120B zero-shot	1.38	0.60	-2.191 (0.103)	1.232 (0.043)
	GPT-OSS-120B few-shot	1.37	0.59	-2.117 (0.102)	1.252 (0.043)
ELA P8	H1	1.13	0.85	0.00*	1.00*
	H2	1.13	0.85	-0.136 (0.070)	0.931 (0.070)
	Claude Sonnet 4.6 zero-shot	0.96	0.62	0.760 (0.079)	3.004 (0.151)
	Claude Sonnet 4.6 few-shot	0.92	0.69	1.009 (0.076)	2.475 (0.127)
	GPT-4.1 zero-shot	1.13	0.52	-0.892 (0.095)	4.086 (0.203)
	GPT-4.1 few-shot	1.19	0.63	-0.901 (0.084)	2.949 (0.150)
	Gemma-4-31B zero-shot	1.12	0.63	-0.394 (0.081)	2.957 (0.150)
	Gemma-4-31B few-shot	1.03	0.64	0.283 (0.079)	2.926 (0.147)
	Granite-3.3-8B zero-shot	1.51	0.83	-1.989 (0.077)	0.039 (0.019)
	Granite-3.3-8B few-shot	1.61	0.73	-2.762 (0.086)	0.222 (0.073)
	Llama-3.3-70B zero-shot	1.07	0.54	-0.211 (0.088)	3.827 (0.189)
	Llama-3.3-70B few-shot	1.15	0.59	-0.806 (0.088)	3.413 (0.172)
	Mistral-Small-3.1 zero-shot	0.75	0.48	3.079 (0.120)	4.435 (0.228)
	Mistral-Small-3.1 few-shot	0.91	0.47	1.273 (0.094)	4.334 (0.213)
	Nemotron-Nano-30B zero-shot	1.33	0.54	-2.717 (0.115)	4.131 (0.213)
	Nemotron-Nano-30B few-shot	1.66	0.56	-4.126 (0.112)	2.398 (0.145)
	Nemotron-Super-120B zero-shot	1.24	0.66	-1.180 (0.083)	2.699 (0.139)
	Nemotron-Super-120B few-shot	1.30	0.70	-1.346 (0.080)	2.142 (0.116)
	GPT-OSS-20B zero-shot	0.82	0.61	1.783 (0.081)	3.063 (0.154)
	GPT-OSS-20B few-shot	0.93	0.65	0.980 (0.078)	2.814 (0.141)
	GPT-OSS-120B zero-shot	0.85	0.59	1.587 (0.082)	3.225 (0.162)
	GPT-OSS-120B few-shot	0.98	0.62	0.606 (0.079)	3.058 (0.153)
ELA P9	H1	1.10	0.76	0.00*	1.00*
	H2	1.10	0.76	-0.151 (0.077)	0.981 (0.039)
	Claude Sonnet 4.6 zero-shot	0.99	0.62	0.641 (0.083)	1.644 (0.056)
	Claude Sonnet 4.6 few-shot	1.06	0.66	0.093 (0.080)	1.474 (0.051)
	GPT-4.1 zero-shot	1.04	0.47	0.064 (0.097)	2.363 (0.077)
	GPT-4.1 few-shot	1.07	0.65	-0.057 (0.081)	1.526 (0.052)
	Gemma-4-31B zero-shot	0.94	0.63	0.999 (0.083)	1.619 (0.055)
	Gemma-4-31B few-shot	1.01	0.65	0.420 (0.082)	1.532 (0.053)
	Granite-3.3-8B zero-shot	1.50	0.77	-2.413 (0.084)	0.332 (0.035)
	Granite-3.3-8B few-shot	1.59	0.76	-2.911 (0.086)	0.041 (0.017)
	Llama-3.3-70B zero-shot	0.99	0.56	0.641 (0.087)	1.914 (0.064)
	Llama-3.3-70B few-shot	1.09	0.62	-0.203 (0.083)	1.637 (0.056)
	Mistral-Small-3.1 zero-shot	0.57	0.52	4.872 (0.161)	2.251 (0.092)
	Mistral-Small-3.1 few-shot	0.76	0.54	2.827 (0.101)	2.086 (0.072)
	Nemotron-Nano-30B zero-shot	1.71	0.55	-4.542 (0.112)	0.904 (0.052)
	Nemotron-Nano-30B few-shot	1.61	0.62	-3.574 (0.096)	0.896 (0.045)
	Nemotron-Super-120B zero-shot	0.97	0.67	0.692 (0.081)	1.436 (0.050)
	Nemotron-Super-120B few-shot	0.99	0.69	0.551 (0.080)	1.343 (0.048)
	GPT-OSS-20B zero-shot	0.88	0.61	1.467 (0.085)	1.689 (0.058)
	GPT-OSS-20B few-shot	0.89	0.64	1.380 (0.083)	1.567 (0.054)
	GPT-OSS-120B zero-shot	0.87	0.57	1.695 (0.088)	1.881 (0.063)
	GPT-OSS-120B few-shot	0.93	0.61	1.112 (0.084)	1.712 (0.058)

Note. H1 is fixed as the reference rater ($\eta = 0$, $\omega = 1$). Values marked with * are anchors, not estimated. Values in parentheses are posterior standard deviations. Raw rating SD is the sample standard deviation.

Table 10: Severity estimates under MFRM and FM-SC, with the shift induced by omitting rater-specific scale use

Prompt	Rater	$\omega_{\text{FM-SC}}$	η_{MFRM}	$\eta_{\text{FM-SC}}$	$\eta\text{-shift}$
Science P1	H1	1.000	0.000	0.000	0.000
	H2	0.997	0.078	0.076	0.001
	Claude Sonnet 4.6 zero-shot	0.795	-0.257	-0.257	0.000
	Claude Sonnet 4.6 few-shot	0.808	0.049	0.053	-0.004
	GPT-4.1 zero-shot	1.006	-0.648	-0.712	0.064
	GPT-4.1 few-shot	1.112	-0.419	-0.487	0.068
	Gemma-4-31B zero-shot	0.772	-0.435	-0.434	-0.002
	Gemma-4-31B few-shot	0.744	-0.648	-0.641	-0.006
	Granite-3.3-8B zero-shot	0.782	-1.218	-1.231	0.013
	Granite-3.3-8B few-shot	1.048	-1.924	-2.135	0.211
	Llama-3.3-70B zero-shot	1.391	0.038	0.025	0.014
	Llama-3.3-70B few-shot	1.324	-0.322	-0.418	0.096
	Mistral-Small-3.1 zero-shot	1.622	0.372	0.514	-0.142
	Mistral-Small-3.1 few-shot	1.606	0.239	0.322	-0.083
	Nemotron-Nano-30B zero-shot	2.073	-0.733	-1.199	0.465
	Nemotron-Nano-30B few-shot	1.293	-2.420	-2.930	0.510
	Nemotron-Super-120B zero-shot	1.392	0.645	0.798	-0.154
	Nemotron-Super-120B few-shot	1.320	0.888	1.073	-0.185
	GPT-OSS-20B zero-shot	0.404	-1.408	-1.285	-0.123
	GPT-OSS-20B few-shot	0.433	-1.341	-1.231	-0.110
	GPT-OSS-120B zero-shot	0.659	-0.577	-0.553	-0.023
	GPT-OSS-120B few-shot	0.741	-0.261	-0.254	-0.007
Science P10	H1	1.000	0.000	0.000	0.000
	H2	1.034	-0.219	-0.222	0.003
	Claude Sonnet 4.6 zero-shot	0.861	-1.154	-1.265	0.111
	Claude Sonnet 4.6 few-shot	0.763	-0.598	-0.664	0.067
	GPT-4.1 zero-shot	1.088	-0.931	-1.092	0.161
	GPT-4.1 few-shot	0.806	-0.575	-0.638	0.063
	Gemma-4-31B zero-shot	1.063	-1.100	-1.292	0.192
	Gemma-4-31B few-shot	1.186	-0.967	-1.187	0.220
	Granite-3.3-8B zero-shot	0.015	-4.609	-3.862	-0.747
	Granite-3.3-8B few-shot	0.017	-4.606	-3.861	-0.745
	Llama-3.3-70B zero-shot	0.930	-2.657	-2.969	0.312
	Llama-3.3-70B few-shot	0.954	-2.127	-2.410	0.283
	Mistral-Small-3.1 zero-shot	1.628	1.456	2.239	-0.783
	Mistral-Small-3.1 few-shot	1.741	0.611	0.989	-0.378
	Nemotron-Nano-30B zero-shot	0.583	0.699	0.613	0.087
	Nemotron-Nano-30B few-shot	0.175	-1.471	-1.372	-0.099
	Nemotron-Super-120B zero-shot	1.304	0.119	0.216	-0.097
	Nemotron-Super-120B few-shot	1.271	-0.443	-0.529	0.086
	GPT-OSS-20B zero-shot	1.135	-1.411	-1.732	0.321
	GPT-OSS-20B few-shot	1.210	-1.408	-1.795	0.386
	GPT-OSS-120B zero-shot	1.232	-1.681	-2.191	0.509
	GPT-OSS-120B few-shot	1.252	-1.610	-2.117	0.507
ELA P8	H1	1.000	0.000	0.000	0.000
	H2	0.931	-0.250	-0.136	-0.115
	Claude Sonnet 4.6 zero-shot	3.004	0.737	0.760	-0.023
	Claude Sonnet 4.6 few-shot	2.475	0.987	1.009	-0.022
	GPT-4.1 zero-shot	4.086	-0.274	-0.892	0.618
	GPT-4.1 few-shot	2.949	-0.615	-0.901	0.287
	Gemma-4-31B zero-shot	2.957	-0.204	-0.394	0.190
	Gemma-4-31B few-shot	2.926	0.342	0.283	0.059
	Granite-3.3-8B zero-shot	0.039	-2.594	-1.989	-0.605
	Granite-3.3-8B few-shot	0.222	-3.369	-2.762	-0.607
	Llama-3.3-70B zero-shot	3.827	0.101	-0.211	0.312
	Llama-3.3-70B few-shot	3.413	-0.405	-0.806	0.401
	Mistral-Small-3.1 zero-shot	4.435	2.001	3.079	-1.078
	Mistral-Small-3.1 few-shot	4.334	1.011	1.273	-0.262
	Nemotron-Nano-30B zero-shot	4.131	-1.454	-2.717	1.262
	Nemotron-Nano-30B few-shot	2.398	-3.713	-4.126	0.412
	Nemotron-Super-120B zero-shot	2.699	-0.929	-1.180	0.251
	Nemotron-Super-120B few-shot	2.142	-1.279	-1.346	0.067
	GPT-OSS-20B zero-shot	3.063	1.575	1.783	-0.208
	GPT-OSS-20B few-shot	2.814	0.932	0.980	-0.048
	GPT-OSS-120B zero-shot	3.225	1.381	1.587	-0.206
	GPT-OSS-120B few-shot	3.058	0.610	0.606	0.004
ELA P9	H1	1.000	0.000	0.000	0.000
	H2	0.981	-0.171	-0.151	-0.021
	Claude Sonnet 4.6 zero-shot	1.644	0.587	0.641	-0.054
	Claude Sonnet 4.6 few-shot	1.474	0.116	0.093	0.023
	GPT-4.1 zero-shot	2.363	0.202	0.064	0.137
	GPT-4.1 few-shot	1.526	-0.012	-0.057	0.045
	Gemma-4-31B zero-shot	1.619	0.896	0.999	-0.103
	Gemma-4-31B few-shot	1.532	0.405	0.420	-0.015
	Granite-3.3-8B zero-shot	0.332	-2.897	-2.413	-0.484
	Granite-3.3-8B few-shot	0.041	-3.572	-2.911	-0.661
	Llama-3.3-70B zero-shot	1.914	0.562	0.641	-0.079
	Llama-3.3-70B few-shot	1.637	-0.114	-0.203	0.088
	Mistral-Small-3.1 zero-shot	2.251	3.377	4.872	-1.495
	Mistral-Small-3.1 few-shot	2.086	2.097	2.827	-0.730
	Nemotron-Nano-30B zero-shot	0.904	-4.725	-4.542	-0.183
	Nemotron-Nano-30B few-shot	0.896	-3.719	-3.574	-0.144
	Nemotron-Super-120B zero-shot	1.436	0.659	0.692	-0.033
	Nemotron-Super-120B few-shot	1.343	0.540	0.551	-0.012
	GPT-OSS-20B zero-shot	1.689	1.276	1.467	-0.191
	GPT-OSS-20B few-shot	1.567	1.250	1.380	-0.131
	GPT-OSS-120B zero-shot	1.881	1.378	1.695	-0.317
	GPT-OSS-120B few-shot	1.712	0.970	1.112	-0.142

Note. $\eta\text{-shift} = \eta_{\text{MFRM}} - \eta_{\text{FM-SC}}$. H1 is the fixed reference rater. Bolded values ($|\eta\text{-shift}| \geq 0.5$) indicate shifts greater than half a logit on the severity scale.