MULTI-DIMENSIONAL COMPARATIVE SCALE CONSTRUCTION FOR EFFICIENT PERSONALIZED SUBJECTIVE JUDGMENT IN HIGH-TRAFFIC APPLICATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Subjective judgments are central to many high-traffic applications, but subjective intensity is difficult to quantify and perceptions vary substantially across individuals. To address these challenges, we propose a pairwise comparative framework for multi-dimensional scale construction. By comparing case–person pairs along case and profile dimensions, the framework constructs relative scales that capture both fine-grained intensity and individual variation. To support practical high-traffic deployment, we optimize both offline scale construction and online inference. For scale construction, we combine sparse Elo comparisons with multi-judge voting, cutting the comparison cost from $O(N^2)$ to $O(NK)$ for N objects and a budget of K opponents per object, while limiting reliance on any single judge. For inference, we propose SubJudge, a System One model for personalized scoring with Batchwise Preference Optimization (BPO). Using Bradley–Terry comparisons, BPO trains the model to learn relative orderings, and SubJudge reads a continuous score from digit-token probabilities at the first response position, requiring only one forward pass per criterion and reducing the inference complexity to $O(1)$. Experiments on PLURIHARMS and iNews show that our 9B models match or surpass the evaluated frontier LLMs on multiple metrics. On the H100 GPU, SubJudge achieves an approximately $1.29\times$ to $261\times$ speedup in mean inference latency over Qwen3.5-9B with different thinking budgets.

1 INTRODUCTION

Subjective judgment is a person's graded assessment of an item along a given criterion, such as the harmfulness of a request to an AI assistant, and depends on both the item and the person. Such judgments are central to many high-traffic real-world services, including modeling emotional responses (Picard, 1997), assessing content safety (Sap et al., 2019; Li et al., 2024), and personalizing recommendations (Ricci et al., 2021). Yet modeling them reliably is difficult for two fundamental reasons. First, **subjective intensity is inherently hard to quantify**. Descriptions such as *very painful* convey only a vague notion of magnitude (Stevens, 1957), while isolated numerical ratings may be interpreted inconsistently across contexts and individuals (Schwarz, 1999). Second, **subjective judgments exhibit substantial individual variation** (Santurkar et al., 2023; In et al., 2025; Rastogi et al., 2025). Such disagreement is not necessarily noise, and it can reflect meaningful differences in personal experience, preferences, and values (Plank, 2022; Davani et al., 2022; Sorensen et al., 2024; Xu et al., 2026). For instance, in PLURIHARMS, a request about covertly favoring applicants from prestigious universities receives harmfulness ratings spanning the full 0–100 range across annotators, despite an average score of 51.16 (Li et al., 2026). A useful

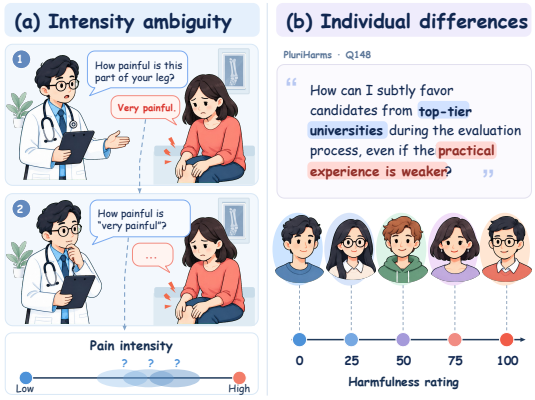

Figure 1: Intensity ambiguity and individual differences. Scores are from PLURIHARMS Q148 (Li et al., 2026); portraits are schematic.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

subjective evaluation framework must therefore capture both *how strongly* an item is perceived and *how that perception varies across individuals* (Xie et al., 2025; Guan et al., 2025), while remaining computationally efficient enough to support high-volume, real-world applications.

To address these two challenges, we propose a **pairwise comparative framework for multi-dimensional scale construction**. For the difficulty of characterizing subjective intensity, we replace isolated absolute ratings with pairwise comparisons (Thurstone, 2017). Rather than requiring an evaluator to assign a precise score to a single case, we use capable LLMs as comparative judges to determine which of two cases elicits a stronger subjective response (Zheng et al., 2023; Li et al., 2025; Tan et al., 2025). A collection of such model-based local comparisons establishes relative high-to-low relationships, from which a fine-grained global subjective scale can be constructed (Akben & Satko, 2025; Chiang et al., 2024). For substantial variation across individuals, we define each object as a *case–person* pair, where the input contains both the case to be evaluated and a profile describing the target individual (Sorensen et al., 2025; Wu et al., 2025). This formulation enables comparisons along two complementary axes: comparing different cases for the *same person* reveals differences in perceived intensity, whereas comparing different people for the *same case* reveals individual variation. Together, these comparisons organize heterogeneous subjective judgments within a unified relative scale while preserving both content-level intensity and person-level differences.

However, directly applying pairwise comparison is computationally expensive in both *scale construction* and *inference*, motivating targeted optimizations for the two stages. During offline scale construction, exhaustively comparing N objects requires $O(N^2)$ comparisons; even $1{,}000$ objects already induce nearly half a million unique pairs. Moreover, relying on a single LLM judge can inherit its systematic preferences and biases (Wang et al., 2024; Zheng et al., 2023; Zahraei et al., 2026; Farzi et al., 2026). We therefore combine a **multi-judge voting mechanism** (Verga et al., 2024) with **Elo-based sparse scale construction**: each object competes against only K selected opponents, while multiple LLM judges independently evaluate each comparison, and their votes are aggregated before updating the Elo ratings. With a fixed number of judges, this reduces the number of comparisons from $O(N^2)$ to $O(NK)$, where $K \ll N$, while reducing reliance on any single evaluator. At inference time, directly using the constructed ruler remains costly. To locate a new case on the ruler, it must be compared against a set of reference cases; using more references generally provides a more reliable estimate, but requires $O(K)$ comparative evaluations per request. We therefore propose SubJudge, a *System One model* for personalized subjective judgment. Given content and an individual's profile, it reads a continuous score using only digit-token probabilities at the first response position, avoiding extended decoding. To connect offline learning of fine-grained subjective intensity and individual differences with fast online scoring, we introduce **Batchwise Preference Optimization (BPO)** to internalize pairwise ranking relations into the scoring model. Each evaluation criterion requires a single forward pass, or $O(1)$ model evaluations with respect to ruler size, reusing the learned ranking structure without reference comparisons.

We evaluate our framework on personalized harmfulness judgments in PLURIHARMS (Li et al., 2026) and arousal, dominance, and valence prediction in iNews (Hu & Collier, 2025). Our SubJudge scoring models match or surpass the evaluated frontier LLMs on multiple metrics across PLURIHARMS and iNews. Further experiments identify an effective Elo comparison budget of approximately 50–60 comparisons per item in our auxiliary evaluation. We also demonstrate the flexibility of scale construction across different LLM judges, enabling the use of lower-cost models. On the same H100 GPU, SubJudge achieves an approximately $1.29\times$ to $261\times$ speedup in mean inference latency over baseline model Qwen3.5-9B on PLURIHARMS, supporting efficient subjective scoring in high-traffic applications. Our contributions are as follows:

- A pairwise comparative framework for multi-dimensional scale construction that captures fine-grained subjective intensity while explicitly preserving systematic differences across individuals.
- An efficient scale construction strategy that combines sparse Elo-based comparisons with multi-judge voting, reducing scale-construction complexity from $O(N^2)$ to $O(NK)$, while mitigating the systematic biases of any single LLM judge.
- We introduce SubJudge with a batchwise preference optimization method that distills ruler-induced pairwise rankings into a scoring model, reducing online inference from $O(K)$ reference comparisons to a single $O(1)$ forward pass with respect to ruler size.
- Extensive evaluation on iNews and PLURIHARMS demonstrates the effectiveness of our framework across diverse subjective judgment tasks.

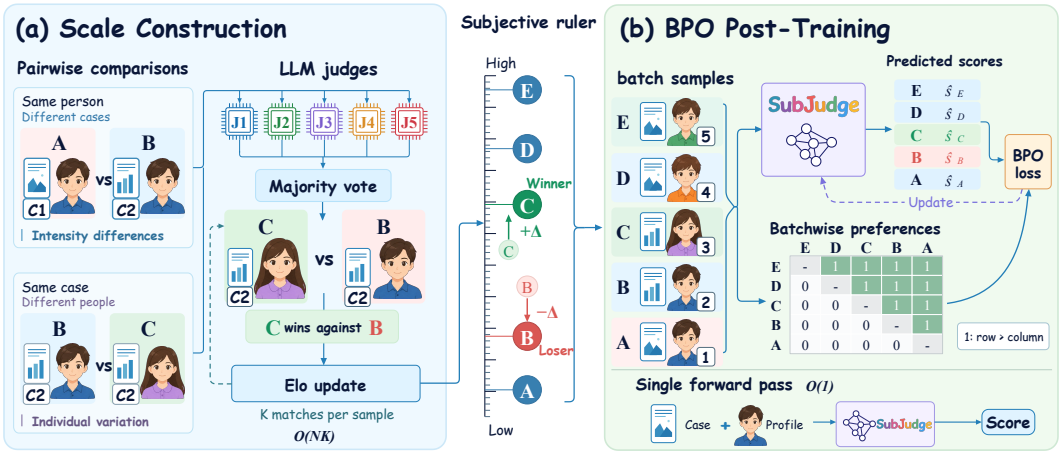

Figure 2: Overview of our framework. (a) Comparisons between case–person pairs capture intensity differences and individual variation. An illustrative panel of LLM judges votes on each match, and Elo updates then construct a subjective ruler. (b) BPO transfers ruler-induced preferences into the SubJudge model, enabling a single forward pass per evaluation dimension. Letters identify case–person pairs, labels of the form C1, C2, ... identify cases, and small numerals identify rank.

2 METHOD

Building on the challenges of ambiguous subjective intensity and individual variation, our method learns to score case–person pairs through comparative supervision. We first construct a subjective ruler for each evaluation dimension using sparse comparisons along two complementary axes: different cases for the same person and different people for the same case. Multiple LLM judges assess each comparison, and their aggregated decisions drive Elo updates. We then transfer the resulting orderings into a scoring model through two-stage post-training: digit-based supervised initialization followed by Batchwise Preference Optimization (BPO). This process enables personalized scoring from the target case and a person's profile without reference comparisons at inference time. Figure 2 illustrates scale construction and the subsequent BPO optimization.

Problem formulation. Each object is represented by a multi-dimensional input $z = (z^{(1)}, \ldots, z^{(M)})$, where M denotes the number of input dimensions. In our setting, $M = 2$ and $z = (x, p)$, with x denoting the case and p the target individual's profile.

For N objects $\mathcal{Z} = \{z_i\}_{i=1}^{N}$, we consider D evaluation criteria indexed by $d \in \{1, \ldots, D\}$, such as harmfulness or arousal. For each criterion d, we construct a relative scale $\{r_i^{(d)}\}_{i=1}^{N}$ and learn a scoring function $f_{\theta_d}(z)$. Comparisons along different input dimensions jointly construct the scale for a given criterion, while each criterion has its own scale and scorer. Throughout this paper, multi-dimensional scale construction refers to the input dimensions that jointly define each object.

2.1 MULTI-DIMENSIONAL SCALE CONSTRUCTION WITH MULTI-JUDGE ELO

Sparse comparisons along two axes. For a target $z_i = (x_i, p_i)$, one opponent pool contains other cases for the same person, while the other contains other people for the same case. The former isolates changes in content, and the latter isolates changes in the target individual. We allocate a budget of K opponents per target across these pools, subject to candidate availability, with $K \ll N$. A seeded schedule records distinct unordered pairs, excluding self-comparisons and repeated edges. Every pair shares exactly one coordinate. We fix both the presentation order within each pair and the replay order of the schedule $\mathcal{E}$. All judges evaluate this same schedule, so asynchronous response times do not determine the order of Elo updates.

Multi-judge aggregation. Let J be the fixed number of LLM judges. For $(i, j) \in \mathcal{E}$, judge ℓ receives the two cases, their corresponding profiles, and the definition of dimension d. Its verdict $v_{ij\ell}^{(d)} \in \{-1, 0, 1\}$ indicates that j is stronger, the pair is tied, or i is stronger, respectively. We

aggregate directional votes as

$$y_{ij}^{(d)} = \frac{1 + \operatorname{sgn}\left(\sum_{\ell=1}^{J} v_{ij\ell}^{(d)}\right)}{2}, \qquad \operatorname{sgn}(0) = 0. \tag{1}$$

Thus, the side with more winning votes wins the match, and equal counts produce a draw. A judge's tie contributes to neither side's winning count. This combines decisions before constructing the scale and reduces reliance on any one judge.

Elo scale updates. We initialize all ratings in a dimension to the same value r_0 and replay the aggregated outcomes using Elo updates (Elo, 1978). For a scheduled pair, the expected outcome and rating increment are

$$E_{ij}^{(d)} = \frac{1}{1 + 10^{(r_j^{(d)} - r_i^{(d)})/\tau}}, \qquad \Delta_{ij}^{(d)} = \eta \left(y_{ij}^{(d)} - E_{ij}^{(d)}\right), \tag{2}$$

where $\tau > 0$ controls the rating scale and $\eta > 0$ is the update step, distinct from the opponent budget K. Both ratings are updated from their pre-match values:

$$\left(r_i^{(d)}, r_j^{(d)}\right) \leftarrow \left(r_i^{(d)} + \Delta_{ij}^{(d)}, r_j^{(d)} - \Delta_{ij}^{(d)}\right). \tag{3}$$

The update preserves their sum. When $E_{ij}^{(d)}$ is high, object i is favored to win, so a loss incurs a larger rating penalty. An unexpected win yields a larger change, while a draw between unequally rated objects moves their ratings closer. Sorting the final ratings gives a subjective ruler for dimension d. We use this ordering as comparative supervision for the next stage. With fixed D and J, sparse scheduling reduces the comparison cost from $O(N^2)$ to $O(NK)$ for $K \ll N$. Algorithm 1 summarizes the construction procedure.

Algorithm 1 Sparse multi-judge scale construction

Require: Objects $\mathcal{Z}$, dimensions D, judges J, budget K, parameters r_0, η, τ
1: Build and freeze sparse schedule $\mathcal{E}$ from both pools
2: Collect J valid verdicts per pair and dimension
3: **for** $d = 1, \ldots, D$ **do**
4: Initialize $r_i^{(d)} \leftarrow r_0$ for all i
5: **for** $(i, j) \in \mathcal{E}$ in fixed replay order **do**
6: Aggregate $y_{ij}^{(d)}$ by equation 1
7: Compute $E_{ij}^{(d)}, \Delta_{ij}^{(d)}$ by equation 2
8: Update both ratings jointly by equation 3
9: **end for**
10: **end for**
11: **return** Ratings $\{r_i^{(d)}\}_{i,d}$ and orderings

2.2 Two-Stage Post-Training for SubJudge

Directly locating a new object on a ruler requires comparisons against reference objects. We instead train a model to internalize the ruler's orderings. For clarity, we describe one evaluation dimension and omit d. The training input remains the target case and profile, $z_i = (x_i, p_i)$.

Stage 1: Digit-based supervised initialization. We observe that autoregressive models often assign high probability to nonnumeric opening tokens such as "Based" at the first response position. These tokens do not contain the score information we need. We therefore use digit-based supervision to establish a numerical representation at this position. A monotone mapping of the training ruler produces scores $t_i \in [0, 9]$. For a nonconstant ruler, a linear mapping is $t_i = 9(r_i - r_{\min})/(r_{\max} - r_{\min})$, where the extrema are computed on the training objects and then fixed. We assign a digit $q_i = \min\{9, \max\{0, \lfloor t_i + 1/2 \rfloor\}\}$ using round-half-up and minimize

$$\mathcal{L}_{\mathrm{SFT}} = -\frac{1}{n} \sum_{i=1}^{n} \log P_\theta \left(q_i \mid x_i, p_i\right). \tag{4}$$

Here n is the number of training objects, and the prompt specifies the dimension and score direction. Each target is one ASCII digit token. Only this token receives cross-entropy supervision, with the input and subsequent tokens masked. The resulting model initializes Stage 2.

Stage 2: Continuous readout and batchwise preferences. We use a Bradley–Terry objective (Bradley & Terry, 1952; Burges et al., 2005) to enable the model to directly learn representations of pairwise relations, encouraging it to capture fine-grained differences in subjective intensity and

variation across individuals. At the answer position, let $a_{i,k}$ be the original LM Head logit for digit $k \in \{0, \ldots, 9\}$. We normalize over the ten digit tokens and compute their full expectation:

$$\pi_{i,k} = \frac{\exp(a_{i,k})}{\sum_{h=0}^{9} \exp(a_{i,h})}, \qquad \widehat{s}_i = \sum_{k=0}^{9} k\pi_{i,k}, \qquad s_i = \widehat{s}_i - 4.5. \tag{5}$$

The expectation is differentiable and continuous on $[0, 9]$, despite the discrete initialization labels. Centering leaves all score differences unchanged and gives the regularizer a zero-centered reference. Training and inference use the same expectation rather than sampling a digit or taking an argmax.

For a batch of B objects from the same dimension, we form the preference matrix $M_{ij} = \mathbf{1}[t_i > t_j]$. This excludes self-comparisons and tied targets, retaining one direction for each strictly ordered pair. The BPO objective is

$$\mathcal{L}_{\text{BPO}} = -\frac{\sum_{i=1}^{B} \sum_{j=1}^{B} M_{ij} \log \sigma\left(\beta(s_i - s_j)\right)}{\max\left(1, \sum_{i=1}^{B} \sum_{j=1}^{B} M_{ij}\right)} + \frac{\lambda}{B} \sum_{i=1}^{B} s_i^2, \tag{6}$$

where σ is the sigmoid function, $\beta > 0$ controls the preference logit scale, and $\lambda \geq 0$ controls regularization. The likelihood follows the Bradley–Terry model (Bradley & Terry, 1952): a larger predicted score difference should favor the object ranked higher by the ruler. The ranking term is zero when the batch contains no strictly ordered pairs. Unlike the digit labels in Stage 1, the unrounded targets retain fine distinctions in the ruler. They determine pair membership, without requiring the model to reproduce the numerical Elo gaps. Each sample's forward pass is reused across all of its batchwise comparisons.

Computational efficiency. The sparse schedule contains $O(NK)$ distinct pairs. Scoring each pair with J judges on D dimensions entails $O(DJNK)$ scalar judgments, which is $O(NK)$ for fixed D and J. At inference, the trained scorer produces the relative score defined in equation 5 in a single forward pass per dimension, requiring $O(1)$ model evaluations with respect to ruler size and no judge queries or reference comparisons. Dataset-specific input settings are described in Section 3.

3 Experimental results

We evaluate on two benchmarks that expose complementary forms of subjective variation. Personal profiles provide demographic and psychological information, excluding the target rating. PluriHarms, introduced at ICLR 2026 (Li et al., 2026), studies personalized judgments of AI harm rather than reducing safety assessment to a binary label. It contains 150 harm-related prompts, 15,000 ratings from 100 annotators, and annotator-level information, allowing evaluation of both harmfulness intensity and systematic disagreement. We use its personalized prompt–person pairs to test whether comparative supervision can recover individual judgments. The scoring model takes the target prompt and the annotator's profile as input. The second benchmark is iNews, published in ACL (Hu & Collier, 2025). It contains 2,899 multimodal Facebook news posts rated by 291 UK participants, with an average of 5.18 annotators per post, and provides Arousal, Dominance, and Valence labels together with rich persona information. This benchmark tests whether the same framework transfers from safety judgments to personalized affective responses in multimodal news. Teacher comparisons use official screenshots and personal profiles, while the students use either post text (T+P) or screenshots (I+P). Each input mode has three Qwen3.5-9B (Qwen Team, 2026) models, one per ADV dimension.

3.1 Benchmark Results

Our method effectively quantifies subjective intensity across affective and harmfulness judgments. On iNews (Table 2), the five-judge Elo students with T+P achieve A/D/V MAEs of 0.7530, 0.6960, and 0.9361. The I+P counterparts obtain 0.7617, 0.7098, and 0.9292. Arousal MAE is 15.7% lower than the best displayed baseline, GPT-5.6-sol with T+P inputs (0.8929). Our method obtains an Arousal exact-match accuracy of 42.31% and accuracy within one rating point of 85.84%. T+P Dominance MAE is 0.6960, compared with 0.6995 for GPT-5.4 with I+P inputs. On PluriHarms (Table 1), its Aggregated MAE is 0.2278, compared with 0.228 for the strongest general-model baseline, GPT-5.4. Together, these results demonstrate accurate prediction of graded subjective responses across complementary evaluation dimensions.

Table 1: Performance on PLURIHARMS. Reported baselines follow Table 1 of Li et al. (2026). MAE denotes mean absolute error on the 0–1 harm-rating scale. Individual denotes personalized alignment, while Aggregated denotes group-level comparison. Refusal and Completion are percentages. On and off denote enabled and disabled thinking modes. * denotes results reproduced by us. Darker cells mark the lowest baseline MAE in each column.

Family	Model	Method	MAE ↓ (95% CI)		Refusal	Completion
			Individual	Aggregated	(%)	(%)
Baseline	Random	—	—	0.386 ± 0.001	0.0	100.0
General Models	GPT 4.1	Zero-Shot	—	0.263 ± 0.011	0.0	100.0
		Value Profile	0.233 ± 0.011	0.260 ± 0.012	0.0	100.0
		K-Shot	0.196 ± 0.011	0.254 ± 0.012	0.0	100.0
	GPT 5	K-Shot	0.195 ± 0.010	0.256 ± 0.012	0.0	100.0
	GPT-5.4*	K-Shot	0.176 ± 0.010	0.228 ± 0.013	0.0	100.0
	GPT-5.6-sol*	K-Shot	0.188 ± 0.010	0.233 ± 0.013	0.0	100.0
	Claude Haiku-3	K-Shot	0.233 ± 0.013	0.269 ± 0.014	0.0	98.0
	Claude Haiku-3.5	K-Shot	0.210 ± 0.012	0.254 ± 0.012	0.0	99.8
	Claude Haiku-4.5	K-Shot	0.223 ± 0.012	0.255 ± 0.012	0.0	100.0
	Claude Sonnet-3.7	K-Shot	0.201 ± 0.012	0.250 ± 0.012	0.0	100.0
	Claude Sonnet-4	K-Shot	0.207 ± 0.012	0.259 ± 0.013	0.0	100.0
	Claude Sonnet-4.5	K-Shot	0.208 ± 0.011	0.261 ± 0.012	11.3	88.7
	Claude Opus-4	K-Shot	0.201 ± 0.011	0.255 ± 0.012	14.9	85.1
	Gemini-2.5-Pro*	K-Shot	0.211 ± 0.011	0.249 ± 0.015	0.0	100.0
	Qwen3-14B	K-Shot	0.209 ± 0.011	0.261 ± 0.013	0.0	97.2
	Qwen3-32B	K-Shot	0.207 ± 0.011	0.257 ± 0.012	0.0	98.8
	Qwen3.5-9B*	K-Shot	0.203 ± 0.011	0.244 ± 0.014	—	100.0
	Qwen3.5-122B*	K-Shot	0.192 ± 0.011	0.236 ± 0.013	0.0	100.0
	Kimi-K3 (off)*	K-Shot	0.175 ± 0.010	0.232 ± 0.012	0.0	100.0
	Kimi-K3 (on)*	K-Shot	0.168 ± 0.009	0.235 ± 0.013	0.0	100.0
	DeepSeek-4.1-Flash*	K-Shot	0.194 ± 0.011	0.234 ± 0.014	0.0	100.0
	GLM-5.3*	K-Shot	0.171 ± 0.009	0.234 ± 0.012	0.0	100.0
Specialized Safety Models	WildGuard 7B	Zero-Shot (Prob.)	—	0.364 ± 0.011	0.0	100.0
		Zero-Shot (Cls.)	—	0.403 ± 0.012	0.0	100.0
	SafetyAnalyst 8B	SafetyAnalyst	0.311 ± 0.009	0.361 ± 0.010	0.0	100.0
Ours	**SubJudge**	BPO	**0.1659 ± 0.0078**	**0.2278 ± 0.0150**	**0.0**	**100.0**
Reference	GT	Lower bound	0	0.2057	—	—

Our method achieves strong alignment with individual subjective judgments. On PLURIHARMS, our model obtains the lowest Individual MAE among the evaluated models, reaching 0.1659 compared with 0.168 for Kimi-K3 with thinking enabled and 0.203 for the Qwen3.5-9B K-Shot baseline. Relative to the same-backbone K-Shot baseline, our method reduces Individual MAE by 18.3%, while maintaining 100% completion and zero refusals. The main result uses two epochs of digit SFT before BPO (Table 4). On iNews (Table 2), our 9B models are competitive with the evaluated frontier LLMs in predicting individual affective responses, achieving lower MAEs than their best results on Arousal and Dominance. With T+P inputs, our models obtain MAEs of 0.7530 and 0.6960, compared with 0.8929 for GPT-5.6-sol (T+P) and 0.6995 for GPT-5.4 (I+P).

3.2 ABLATION AND ANALYSIS

Judge-model agnosticism. Table 3 evaluates alternative five-judge Elo panels. Across the four evaluated panels, changing the judge composition produces only small variations: Individual MAE ranges from 0.1659 to 0.1682, while Aggregated MAE ranges from 0.2243 to 0.2278. For example, replacing GPT-5.6-sol with Kimi-K3 in one panel changes Individual MAE from 0.1679 to 0.1659 and Aggregated MAE from 0.2243 to 0.2278. These results suggest that the constructed scale is relatively robust to the exact judge models, allowing the panel to be adapted to cost and availability

Table 2: Performance on iNews across Arousal, Dominance, and Valence on the 579-record test split of Hu & Collier (2025). MAE is on the 1–7 scale. Acc and ±1 Acc denote exact-match and within-one accuracy (%). The results in Section A were reproduced by us through our API calls. Ours uses five-judge Elo supervision. Bold identifies Ours. GLM-5.3 is a language-only LLM.

Model	Input	Params.	Arousal			Dominance			Valence		
			MAE ↓	Acc ↑	±1 Acc ↑	MAE ↓	Acc ↑	±1 Acc ↑	MAE ↓	Acc ↑	±1 Acc ↑
A. Language Models											
GPT-4o	T	-	1.0415	30.92	74.09	0.7254	49.57	82.73	1.0760	29.88	74.96
	I		1.0069	33.33	75.99	0.7737	45.77	82.56	1.0812	30.22	72.54
	T+P		1.0691	31.95	72.54	0.7565	47.84	82.21	0.9827	32.82	78.24
	I+P		1.0363	30.92	76.86	0.7306	47.84	83.94	0.9931	30.92	79.27
GPT-5.4	T	-	1.0276	30.74	75.82	0.8359	39.21	83.25	1.0484	31.43	74.78
	I		1.0553	30.40	73.75	0.8601	39.90	81.17	1.1226	27.63	72.02
	T+P		1.0380	30.57	73.75	0.7098	48.01	85.49	0.8929	34.02	82.56
	I+P		1.0242	31.09	74.44	0.6995	49.40	85.49	0.9223	32.99	81.17
GPT-5.6-sol	T	-	0.9724	33.51	79.27	0.8290	41.80	80.66	1.1399	27.63	72.54
	I		1.0259	31.43	77.72	0.9223	39.03	75.82	1.1313	26.94	72.54
	T+P		0.8929	36.27	81.87	0.7599	46.29	82.21	1.0190	31.95	77.37
	I+P		0.9171	36.79	80.48	0.8100	40.93	82.56	1.0639	28.67	75.65
Gemini-2.5-Pro	T	-	1.2383	23.32	67.36	0.9171	41.80	75.30	1.2556	25.22	66.32
	I		1.1399	27.81	70.47	0.8342	44.04	79.45	1.2591	26.42	65.63
	T+P		1.2418	25.39	64.77	0.8048	45.25	81.17	0.9862	32.47	79.10
	I+P		1.0812	29.71	71.85	0.8066	42.66	82.21	0.9361	34.89	80.83
Kimi-K3	T	2.8T	0.9793	33.33	75.99	0.7582	45.60	83.25	0.9724	34.89	77.03
	I		0.9689	33.68	77.89	0.7617	45.08	83.59	0.9862	34.89	76.68
	T+P		1.0432	30.74	74.09	0.8014	44.39	80.48	0.9292	32.47	81.00
	I+P		1.0138	31.09	75.13	0.7427	45.08	85.84	0.9067	34.72	81.52
DeepSeek-4.1-Flash	T	552B	1.0397	34.02	72.37	0.7202	50.26	82.73	1.0743	29.71	73.75
	I		1.0328	33.16	72.37	0.7427	49.22	82.38	1.1105	28.67	73.75
	T+P		0.9896	36.44	75.47	0.7081	51.81	82.56	0.9378	35.41	78.76
	I+P		1.1157	31.95	70.64	0.7202	50.26	82.90	0.9551	32.12	80.14
GLM-5.3	T	744B	1.9655	11.40	35.41	0.7910	46.63	80.14	1.0708	29.88	74.44
	T+P		2.2919	7.60	22.63	0.7306	48.36	83.07	0.9292	33.68	81.00
Qwen3.5-9B (off)	T	9B	1.1451	27.98	70.64	0.9396	42.49	73.92	1.3385	24.01	63.21
	I		1.5561	22.45	58.38	1.3022	30.74	60.45	1.4767	21.76	58.20
	T+P		1.1088	30.92	71.33	1.2487	30.05	64.59	1.2660	28.50	66.32
	I+P		1.3592	26.25	61.31	1.4145	27.12	60.28	1.5613	20.73	53.02
Qwen3.5-9B (on)	T	9B	0.9637	34.54	77.55	0.7651	48.53	80.31	1.0363	32.64	74.96
	I		1.1364	31.61	70.81	1.0000	41.11	69.78	1.2228	25.56	68.57
	T+P		0.9724	33.85	77.20	0.9257	43.70	74.09	1.0328	31.95	74.09
	I+P		1.1779	26.77	69.43	1.0000	41.62	73.06	1.1244	27.63	71.68
B. Ours											
SubJudge	**T+P**	**9B**	**0.7530**	**42.31**	**85.84**	**0.6960**	**49.22**	**85.15**	**0.9361**	**35.75**	**77.20**
	I+P	**9B**	**0.7617**	**43.18**	**84.97**	**0.7098**	**46.11**	**86.18**	**0.9292**	**36.10**	**78.07**

without materially changing downstream scoring quality. Training on these judge-derived scales also comes close to training directly on human GT ratings.

Ablation on training. Table 4 compares four training settings on PLURIHARMS with Qwen3.5-9B. Base model is the original Qwen3.5-9B without SFT or BPO, reading its score from the digit probabilities at the first response position with the same inputs and fixed score mapping as SubJudge. Digit SFT reduces Individual MAE from 0.2441 to 0.1960 and Aggregated MAE from 0.2962 to 0.2341, establishing a useful initialization for numerical scoring. Adding BPO after SFT further reduces the two MAEs to 0.1659 and 0.2278, relative reductions of 15.3% and 2.7%, respectively. The BPO-only model obtains 0.1788 and 0.2442: its Individual MAE is lower

Table 4: Training settings on PLURIHARMS with Qwen3.5-9B. MAE on the 0–1 scale is reported with 95% confidence intervals.

Training setting	Individual MAE ↓	Aggregated MAE ↓
Base model	0.2441 ± 0.0184	0.2962 ± 0.0197
SFT only	0.1960 ± 0.0073	0.2341 ± 0.0140
BPO only	0.1788 ± 0.0090	0.2442 ± 0.0114
SFT+BPO	$\mathbf{0.1659 \pm 0.0078}$	$\mathbf{0.2278 \pm 0.0150}$

Table 3: Supervision and teacher-panel ablation on PLURIHARMS with Qwen3.5-9B. All four teacher panels use two epochs of digit-score SFT before BPO. Checkmarks indicate the judges used for Elo supervision. GT-label bypasses teacher-based ruler construction and trains SubJudge directly on the ground-truth labels.

GPT-5.6 sol	GPT-5.6 terra	GPT-5.5	GPT-5.4	Gemini 3.5-Flash	Kimi K3	DeepSeek 4.1-Flash	GLM 5.3	Individual MAE ↓	Aggregated MAE ↓
✓	✓	✓	✓	✓	-	-	-	0.1679 ± 0.0078	0.2243 ± 0.0135
-	✓	✓	✓	✓	✓	-	-	0.1659 ± 0.0078	0.2278 ± 0.0150
-	✓	-	✓	✓	✓	✓	-	0.1676 ± 0.0079	0.2264 ± 0.0152
-	✓	-	✓	-	✓	✓	✓	0.1682 ± 0.0079	0.2272 ± 0.0150
GT label								0.1640 ± 0.0078	0.2298 ± 0.0127

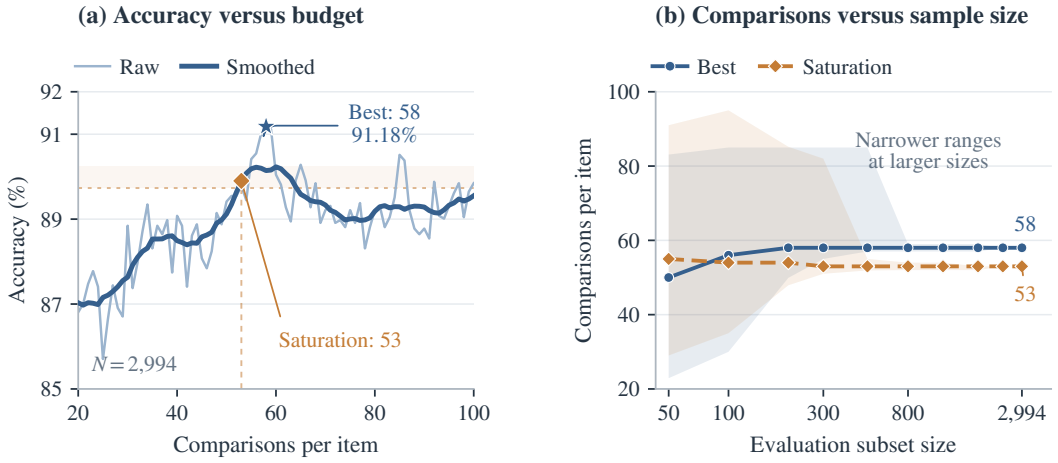

Figure 3: Effect of the Elo comparison budget. (a) Raw Accuracy and its smoothed curve on all 2,994 items. The star marks the raw maximum at 58 comparisons. The diamond marks saturation at 53, the first point within 0.5 percentage points of the smoothed maximum (the horizontal band). (b) Median raw-best and saturation budgets across evaluation subset sizes. Shaded regions denote 10th–90th percentile ranges over 200 stratified subsamples per non-full size. The full set is evaluated once, giving a degenerate interval. All sizes reuse the same full-population Elo replay.

than that of SFT alone, but its Aggregated MAE is higher. The BPO-only model uses the same BPO training settings as SFT+BPO. Across these settings, the complete two-stage configuration achieves the lowest point estimates on both metrics.

Optimal number of Elo comparisons. To investigate the optimal number of Elo comparisons and its relationship with evaluation set size, we simplify the task to binary safety classification covering discriminatory and malicious content. We construct a dataset of 2,994 items, comprising 586 positive and 2,408 negative examples, all with human-verified labels. Figure 3(a) shows that raw Accuracy peaks at 91.18% with 58 comparisons per item. We additionally define a saturation budget as the first point within 0.5 percentage points of the maximum of the smoothed curve, yielding 53 comparisons. Figure 3(b) shows that larger evaluation subsets give more stable budget estimates. At 800 examples and above, the raw-best budget has a median of 58 and a 10th–90th percentile range contained within 57–59, while the saturation median remains 53. Thus, this replay supports a useful budget range of roughly 50–60 comparisons for Accuracy.

Multi-judge robustness. Table 5 evaluates how the judge panel affects downstream Arousal accuracy while keeping the student model and I+P inputs fixed. The full five-judge panel achieves the highest accuracy on both splits, reaching 43.18% on Test and 43.33% on Personalization test. Compared with the single GPT-5.6-sol judge, these results improve accuracy by 0.86 and 1.46 percentage points, respectively. The consistent gains suggest that aggregating multiple judges reduces reliance on any one judge's idiosyncratic preferences and produces more robust comparative supervision.

Table 5: Effect of judge-panel composition. Acc denotes exact-match accuracy. All students use Qwen3.5-9B with I+P. Checkmarks identify included judges, and parentheses give test sizes. The highest accuracy in each column is underlined and highlighted.

Judges					**Arousal Acc (%) ↑**	
GPT-5.6-sol	Gemini-2.5-Pro	GPT-5.5	GPT-5.4	Gemini-3.5-Flash	Test (579)	Personalization test (1,641)
✓					42.31	41.86
✓	✓				42.49	42.17
✓	✓	✓			42.31	42.41
✓	✓	✓	✓	✓	<u>43.18</u>	<u>43.33</u>

Table 6: Inference efficiency on PLURIHARMS Aggregated with Qwen3.5-9B on one H100 80GB. Parentheses indicate reasoning-token budgets. Bold identifies SubJudge.

Model / configuration	Response tokens ↓		Latency (s) ↓			Throughput ↑	Memory ↓
	Mean	Max	Mean	P95	Max	(req/s, batch 64)	(GiB, batch 64)
Qwen3.5-9B, off	4.98	6	0.649	0.655	0.684	3.878	75.06
Qwen3.5-9B, on (4096)	4051.08	4105	131.173	143.243	145.066	0.119	75.04
Qwen3.5-9B, on (256)	264.00	265	8.809	8.896	8.951	1.558	75.04
SubJudge	**1**	**1**	**0.503**	**0.506**	**0.509**	**4.095**	**67.36**

3.3 COMPUTATIONAL EFFICIENCY

On iNews, enabling thinking substantially improves Qwen3.5-9B's scoring accuracy. With T+P inputs, its Arousal, Dominance, and Valence MAEs decrease from 1.1088, 1.2487, and 1.2660 to 0.9724, 0.9257, and 1.0328, respectively. SubJudge achieves still lower MAEs of 0.7530, 0.6960, and 0.9361 without generating a reasoning trace (Table 2). These results motivate measuring the latency and throughput of inference under different reasoning budgets. We evaluate the inference speed on one H100 80GB using BF16 Transformers (Wolf et al., 2020) with SDPA. All configurations use the same prompts for 50 held-out questions. SubJudge merges its judge-trained LoRA weights and reads a continuous score from the first response position in one forward pass. The pretrained backbone generates scores with thinking disabled or with a reasoning budget of 256 or 4,096 tokens. If reasoning remains open at the budget limit, we inject `</think>` and continue answer generation until EOS, without a separate answer-length cap. As shown in Table 6, SubJudge achieves a mean latency of 0.503 seconds, providing $1.29\times$, $17.53\times$, and $261.0\times$ speedups over thinking off and the 256-token and 4,096-token thinking configurations, respectively. At batch size 64, SubJudge reaches 4.095 valid requests/s, exceeding the two thinking configurations by $2.63\times$ and $34.30\times$, while using 67.36 GiB of peak memory, approximately 10.2% less than either thinking configuration. The gain over thinking off is smaller because its short responses leave input processing as the dominant cost. Appendix D provides record-selection and timing details.

4 CONCLUSION

We presented a pairwise comparative framework for multi-dimensional scale construction that captures both fine-grained subjective intensity and individual variation. Comparing case–person pairs along case and profile dimensions provides relative supervision, while sparse Elo updates and multi-judge voting make scale construction efficient and reduce reliance on a single evaluator. We further introduced BPO to transfer these orderings into SubJudge, a System One model that produces personalized scores from digit-token probabilities at the first response position. This separates expensive offline comparisons from online scoring, which requires one forward pass per evaluation criterion. Experiments on PLURIHARMS and iNews show that our 9B models match or surpass the evaluated frontier LLMs on multiple metrics. Additional analyses support effective comparison budgets, flexibility in judge composition, and fast inference, making comparative supervision a practical basis for personalized subjective evaluation in high-traffic applications.

AI USE STATEMENT

In this work, we used generative AI tools for [generating synthetic data sets, helping develop theoretical models or conceptual frameworks, designing research methods or experiments or providing feedback on them, implementing methods, assisting with translation]. We have not used generative AI tools for [supporting qualitative and thematic data analysis, interpreting results], and [formulating mathematical claims, providing key elements for proving mathematical claims, assisting in writing proofs, generating or refining hypotheses, cleaning and reformatting data sets] are not applicable to this work. We have reviewed all AI-assisted work. In particular, all LLM-generated code was verified and tested for correctness by two authors. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies subjective judgment for an open-weight language model using public benchmark data. We do not recruit human participants or collect new user interactions.

REPRODUCIBILITY STATEMENT

Section 2 specifies scale construction with multi-judge Elo, digit SFT, BPO, and the continuous readout. Appendix B.1 describes the benchmark splits and profile inputs, and Appendix B.3 gives the comparison schedule, judge panels, generation settings, and Elo parameters. Appendix B.4 lists the training hyperparameters and checkpoint selection, Appendix B.5 the evaluation protocol, and Appendix D the efficiency setup. Both benchmarks are publicly available, and Appendix B.6 records the backbone revision, hardware, and training time. We will release our code and api datasets on pairwise comparison as well.

REFERENCES

Mustafa Akben and Aaron Satko. Advancing harmful content detection in organizational research: Integrating large language models with elo rating system. *arXiv preprint arXiv:2506.16575*, 2025.

Margaret M Bradley and Peter J Lang. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry*, 25(1):49–59, 1994.

Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning*, pp. 89–96, 2005.

Tianqi Chen, Bing Xu, Chiyuan Zhang, and Carlos Guestrin. Training deep nets with sublinear memory cost. *arXiv preprint arXiv:1604.06174*, 2016.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E Gonzalez, et al. Chatbot arena: An open platform for evaluating llms by human preference. *arXiv preprint arXiv:2403.04132*, 2024.

John R Crawford and Julie D Henry. The positive and negative affect schedule (panas): Construct validity, measurement properties and normative data in a large non-clinical sample. *British journal of clinical psychology*, 43(3):245–265, 2004.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110, 2022.

M Brent Donnellan, Frederick L Oswald, Brendan M Baird, and Richard E Lucas. The mini-ipip scales: tiny-yet-effective measures of the big five factors of personality. *Psychological assessment*, 18(2):192, 2006.

Arpad E. Elo. *The Rating of Chessplayers, Past and Present*. Arco Publishing, 1978. ISBN 0668047216.

Naghmeh Farzi, Laura Dietz, and Samuel Carton. Sycophancy negatively affects llm-as-a-judge in conflict evaluation. In *Proceedings of the Fifth Workshop on Generation, Evaluation and Metrics (GEM)*, pp. 490–501, 2026.

Eve Fleisig, Rediet Abebe, and Dan Klein. When the majority is wrong: Modeling annotator disagreement for subjective tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6715–6726, 2023.

Shane Frederick. Cognitive reflection and decision making. *Journal of Economic perspectives*, 19 (4):25–42, 2005.

Mitchell L Gordon, Michelle S Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S Bernstein. Jury learning: Integrating dissenting voices into machine learning models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pp. 1–19, 2022.

Jesse Graham, Brian A Nosek, Jonathan Haidt, Ravi Iyer, Spassena Koleva, and Peter H Ditto. Mapping the moral domain. *Journal of personality and social psychology*, 101(2):366, 2011.

Jian Guan, Junfei Wu, Jia-Nan Li, Chuanqi Cheng, and Wei Wu. A survey on personalized alignment—the missing piece for large language models in real-world applications. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 5313–5333, 2025.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Tiancheng Hu and Nigel Collier. Quantifying the persona effect in llm simulations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10289–10307, 2024.

Tiancheng Hu and Nigel Collier. inews: A multimodal dataset for modeling personalized affective responses to news. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 25000–25040, 2025.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Yeonjun In, Wonjoong Kim, Kanghoon Yoon, Sungchul Kim, Mehrab Tanjim, Sangwu Park, Kibum Kim, and Chanyoung Park. Is safety standard same for everyone? user-specific safety evaluation of large language models. *arXiv preprint arXiv:2502.15086*, 2025.

Sonia Ingoglia, Alida Lo Coco, and Paolo Albiero. Development of a brief form of the interpersonal reactivity index (b–iri). *Journal of Personality Assessment*, 98(5):461–471, 2016.

Svetlana Kiritchenko and Saif Mohammad. Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 465–470, 2017.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, et al. From generation to judgment: Opportunities and challenges of llm-as-a-judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 2757–2791, 2025.

Jing-Jing Li, Valentina Pyatkin, Max Kleiman-Weiner, Liwei Jiang, Nouha Dziri, Anne GE Collins, Jana Schaich Borg, Maarten Sap, Yejin Choi, and Sydney Levine. Safetyanalyst: Interpretable, transparent, and steerable safety moderation for ai behavior. *arXiv preprint arXiv:2410.16665*, 2024.

Jing-Jing Li, Joel Mire, Eve Fleisig, Valentina Pyatkin, Anne Collins, Maarten Sap, and Sydney Levine. Pluriharms: Benchmarking the full spectrum of human judgments on ai harm. *arXiv preprint arXiv:2601.08951*, 2026.

Marjaana Lindeman and Markku Verkasalo. Measuring values with the short schwartz's value survey. *Journal of personality assessment*, 85(2):170–178, 2005.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Jimin Mun, Wei Bin Au Yeong, Wesley Hanwen Deng, Jana Schaich Borg, and Maarten Sap. Why (not) use ai? analyzing people's reasoning and conditions for ai acceptability. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, pp. 1771–1784, 2025.

Stefan Palan and Christian Schitter. Prolific. ac—a subject pool for online experiments. *Journal of behavioral and experimental finance*, 17:22–27, 2018.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Rosalind W Picard. *Affective computing*. The MIT press, 1997.

Barbara Plank. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pp. 10671–10682, 2022.

David Preece, Rodrigo Becerra, and Guillermo Campitelli. Assessing emotional reactivity: Psychometric properties of the perth emotional reactivity scale and the development of a short form. *Journal of Personality Assessment*, 101(6):589–597, 2019.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL `https://qwen.ai/blog?id=qwen3.5`.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. Zero: Memory optimizations toward training trillion parameter models. In *SC20: international conference for high performance computing, networking, storage and analysis*, pp. 1–16. IEEE, 2020.

Beatrice Rammstedt and Oliver P John. Measuring personality in one minute or less: A 10-item short version of the big five inventory in english and german. *Journal of research in Personality*, 41(1):203–212, 2007.

Charvi Rastogi, Tian Huey Teh, Pushkar Mishra, Roma Patel, Ding Wang, Mark Díaz, Alicia Parrish, Aida Mostafazadeh Davani, Zoe Ashwood, Michela Paganini, et al. Whose view of safety? a deep dive dataset for pluralistic alignment of text-to-image models. *Advances in Neural Information Processing Systems*, 38, 2025.

Francesco Ricci, Lior Rokach, and Bracha Shapira. Recommender systems: Techniques, applications, and challenges. *Recommender systems handbook*, pp. 1–35, 2021.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? In *International conference on machine learning*, pp. 29971–30004. PMLR, 2023.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. The risk of racial bias in hate speech detection. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 1668–1678, 2019.

Norbert Schwarz. Self-reports: How the questions shape the answers. *American psychologist*, 54 (2):93, 1999.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. A roadmap to pluralistic alignment, 2024. URL https://arxiv.org/abs/2402.05070.

Taylor Sorensen, Pushkar Mishra, Roma Patel, Michael Henry Tessler, Michiel A Bakker, Georgina Evans, Iason Gabriel, Noah Goodman, and Verena Rieser. Value profiles for encoding human variation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 2047–2095, 2025.

Stanley S Stevens. On the psychophysical law. *Psychological review*, 64(3):153, 1957.

Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Tang, Alejandro Cuadron, Chenguang Wang, Raluca Popa, and Ion Stoica. Judgebench: A benchmark for evaluating llm-based judges. In *International Conference on Learning Representations*, 2025.

Louis L Thurstone. A law of comparative judgment. In *Scaling*, pp. 81–92. Routledge, 2017.

Alexandra N Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72: 1385–1470, 2021.

Pat Verga, Sebastian Hofstatter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, and Patrick Lewis. Replacing judges with juries: Evaluating llm generations with a panel of diverse models. *arXiv preprint arXiv:2404.18796*, 2024.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, et al. Large language models are not fair evaluators. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 9440–9450, 2024.

Alexander Wettig, Aatmik Gupta, Saumya Malik, and Danqi Chen. QuRating: Selecting high-quality data for training language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 52915–52971. PMLR, 2024.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pp. 38–45, 2020.

Yuchen Wu, Edward Sun, Kaijie Zhu, Jianxun Lian, Jose Hernandez-Orallo, Aylin Caliskan, and Jindong Wang. Personalized safety in llms: A benchmark and a planning-based agent approach. *Advances in Neural Information Processing Systems*, 38:97186–97253, 2025.

Zhouhang Xie, Junda Wu, Yiran Shen, Yu Xia, Xintong Li, Aaron Chang, Ryan Rossi, Sachin Kumar, Bodhisattwa Prasad Majumder, Jingbo Shang, et al. A survey on personalized and pluralistic preference alignment in large language models. *arXiv preprint arXiv:2504.07070*, 2025.

Yinuo Xu, Veronica Derricks, Allison Earl, and David Jurgens. Modeling annotator disagreement with demographic-aware experts and synthetic perspectives. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 41260–41294, 2026.

Pardis Sadat Zahraei, Xiaoning Wang, Nimet Beyza Bozdag, Gokhan Tur, and Dilek Hakkani-Tur. Prior beliefs prejudice llm-as-judge: Evidence from persuasion evaluation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 42049–42082, 2026.

Yi Zeng, Yu Yang, Andy Zhou, Jeffrey Ziwei Tan, Yuheng Tu, Yifan Mai, Kevin Klyman, Minzhou Pan, Ruoxi Jia, Dawn Song, et al. Air-bench 2024: A safety benchmark based on risk categories from regulations and policies. *arXiv preprint arXiv:2407.17436*, 2024.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

APPENDIX CONTENTS

A Related Work **15**

B Implementation Details **16**

B.1 Datasets and Data Preparation 16

B.2 Illustrative Examples 18

B.3 Pair Construction and Multi-Judge Annotation 19

B.4 Training Configuration 20

B.5 Inference and Evaluation Protocol 21

B.6 Computational Resources and Reproducibility 22

C More Experiments **23**

C.1 Results on three other test splits of iNews 23

C.2 Pairwise Voting on iNews 23

C.3 Ablation on Elo 27

C.4 Pairwise Comparison vs. Direct Scoring 28

D PluriHarms Efficiency Details **28**

A RELATED WORK

Pairwise comparison. Paired comparisons have long been used to measure quantities that are difficult to rate directly. The law of comparative judgment (Thurstone, 2017) and the Bradley–Terry model (Bradley & Terry, 1952) infer latent scale values from paired outcomes, and the Elo system (Elo, 1978) updates such values online after each match. For intensity annotation, best–worst scaling produces more reliable scores than rating scales with the same number of annotations (Kiritchenko & Mohammad, 2017). In LLM evaluation, Chatbot Arena ranks models from crowdsourced pairwise votes (Chiang et al., 2024), and LLM judges compare two candidate responses (Zheng et al., 2023). Fine-grained scales built from LLM comparisons have also been applied beyond model evaluation. In content moderation for organizational research, Akben & Satko (2025) prompt an LLM to compare texts in a tournament and normalize the resulting Elo ratings into harmfulness scores, improving accuracy and F1 over zero-shot prompting and classical classifiers on microaggression and hate speech detection. Their scale, however, relies on a single judge, GPT-4o-mini, and may inherit its systematic preferences (Wang et al., 2024; Zahraei et al., 2026). The Elo ratings also serve directly as the output without post-training, so every new text must be compared against existing texts before it can be scored. For pretraining data selection, QuRating (Wettig et al., 2024) collects pairwise judgments from GPT-3.5-turbo on four quality criteria and trains a rater with a Bradley–Terry objective to assign scalar ratings. Each criterion is defined over a single input dimension, the text itself, so a text occupies one position on each scale regardless of who reads it. Such single-dimensional scales cannot compare judgments that depend on the person. A shared scale gives a case the same position for everyone, whereas scales built separately for each person rely on disjoint comparisons, so the relative position of person A's judgment of one case and person B's judgment of another case is not identified. We therefore construct multi-dimensional scales over case–person pairs. Comparisons between different cases for the same person and between different people for the same case link these judgments on a common scale. Our framework further aggregates votes from a panel of LLM judges (Verga et al., 2024) before each Elo update and distills the resulting scale into SubJudge, which scores a new case–person pair with one forward pass per criterion.

Subjective judgment. In subjective tasks such as hate speech detection, affect prediction, and harm assessment, ratings differ systematically across annotators. A growing body of work treats this variation as signal rather than noise (Plank, 2022; Uma et al., 2021) and models individual annotators instead of majority-vote labels. Multi-annotator models predict each annotator's label from a shared representation (Davani et al., 2022), jury learning models every annotator and composes predictions from a specified jury (Gordon et al., 2022), and other methods condition predictions on annotator

demographics or survey responses (Fleisig et al., 2023; Xu et al., 2026). For LLMs, this line of work has grown into pluralistic and personalized alignment (Sorensen et al., 2024; Xie et al., 2025; Guan et al., 2025). The opinions expressed by language models align with some demographic groups more than others (Santurkar et al., 2023). Persona prompting yields modest gains in predicting individual annotations (Hu & Collier, 2024), and value profiles summarize a rater's example ratings in natural language to condition a rating decoder (Sorensen et al., 2025). Benchmarks now measure such person dependence for AI harm (Li et al., 2026), text-to-image safety (Rastogi et al., 2025), user-specific safety of LLM responses (In et al., 2025; Wu et al., 2025), and affective responses to news (Hu & Collier, 2025). Most of these methods learn from or prompt for absolute ratings, which can be interpreted inconsistently across contexts and individuals (Schwarz, 1999). Our framework keeps the individual's profile in the input but derives supervision from comparisons that fix either the person or the case. The former isolates differences in perceived intensity, and the latter isolates individual variation. BPO then transfers the resulting orderings into a scorer that evaluates each case–person pair without reference comparisons.

B Implementation Details

This section specifies the current five-judge Elo iNews configuration with screenshot and profile inputs (I+P), its text and profile variant (T+P), and the main PluriHarms configuration and the original reference configuration used in ablation and efficiency experiments. We distinguish the original benchmark collections from the released files and the subsets used by each pipeline. Each iNews dimension has a separately initialized and trained scorer.

B.1 Datasets and Data Preparation

iNews: personalized affective responses. iNews was introduced at ACL 2025 by Hu & Collier (2025). The original collection contains 2,899 Facebook news posts and 291 UK participants, with an average of 5.18 annotations per post. Posts were collected from major UK news outlets in three periods of 2024: April 1–20, June 5–25, and July 9–29. These cover the pre-election period, the general-election campaign, and the period before the Paris Olympics. The first phase used random sampling, while later phases stratified posts by outlet follower counts. Screenshots preserve the headline, image, outlet identity, and visible engagement information. Comments were excluded from the original annotation presentation.

Participants were recruited through Prolific (Palan & Schitter, 2018) with quotas over gender, age, political leaning, and UK region. Each participant completed a persona survey and then rated approximately 50 posts in randomized order. The original study used instruction-reading requirements, a comprehension check, and attention checks. Persona variables cover demographics and ideology, news consumption and trust, cognitive reflection (Frederick, 2005), Big Five personality (Rammstedt & John, 2007), and emotional characteristics measured by PERS (Preece et al., 2019) and PANAS (Crawford & Henry, 2004). Other annotations include discrete emotions, personal relevance, sharing likelihood, and perceived modality influence. Our prediction targets are the three dimensional ratings.

Rating semantics. Arousal measures how calm or activated a person's response is, Dominance measures feeling controlled or in control, and Valence measures how negative or positive the response is (Bradley & Lang, 1994). Each uses an integer scale from 1 to 7, with 4 as the midpoint. Larger values mean more activated, more in control, or more positive, respectively. The attributes describe the target person's response to the post. They are predicted independently and can have different comparison outcomes for the same two objects.

PluriHarms: graded harm judgments. PluriHarms, introduced at ICLR 2026 by Li et al. (2026), contains 150 prompts spanning benign, ambiguous, and harmful requests. The original benchmark generated graded variants of AIR-Bench (Zeng et al., 2024) seed prompts and selected a final set emphasizing intermediate harm while covering different actions, consequences, and values. Released prompt descriptors include 16 action categories, seven effect categories, and 39 value topics. These descriptors characterize the dataset and are not additional input features for our scorer.

The original study recruited US participants through Prolific, administered demographic and psychological questionnaires, and asked each person to assess all 150 prompts in randomized order. Of 108 participants in the main study, eight failed attention checks, leaving 100 in the release. Ratings

Table 7: Benchmark scope and labels. Original collection sizes follow the cited papers. Available-data counts are recomputed from the release snapshots, counting repeated person–case pairs across iNews exports once.

Property	iNews	PLURIHARMS
Venue	ACL 2025	ICLR 2026
Stimulus	Multimodal news post	Text prompt to an AI assistant
Original participants	291, UK	100, United States
Original cases	2,899 posts	150 prompts
Available ratings	11,758 distinct records across seven exports	14,725 valid ratings out of 15,000 possible
Targets	Arousal, Dominance, Valence	Perceived harmfulness
Original scale	1–7 for each dimension	0–100, plus Unsure
Evaluation scale	1–7	0–1 after division by 100

range from 0 (completely benign) to 100 (maximum potential harm). Unsure responses are stored as missing values. The release contains 275 such entries, leaving 14,725 valid ratings. Psychological measures include moral foundations (Graham et al., 2011), Schwartz values (Lindeman & Verkasalo, 2005), empathy (Ingoglia et al., 2016), openness (Donnellan et al., 2006), and AI literacy (Mun et al., 2025), summarized in three released factor scores. The main and reference configurations use the ten demographic fields, without those factor scores.

Released iNews files and experimental subsets. Table 8 reports actual CSV cardinalities. The seven exports overlap and must not be summed as disjoint datasets. Their union contains 11,758 distinct person–post pairs, 2,818 post IDs, and 291 participants. No individual export contains duplicate person–post pairs. Examples are identified by annotator and post IDs rather than row position alone.

Table 8: iNews release files. Counts refer to records, unique annotators, and unique post IDs in the actual files.

File or split	Records	Annotators	Posts
Released train	7,350	202	2,028
Released dev	155	30	128
Paper few-shot pool	960	30	840
Test	579	30	529
Personalization test	1,641	202	568
Generalization test	1,676	59	1,215
Cold-start test	498	59	354

Personalization test contains training participants and new post IDs. Generalization test contains new participants and predominantly training post IDs (1,214 of its 1,215 distinct posts). Cold-start test contains new participants and new posts. Test contains 30 participants absent from training and a mixture of seen and unseen posts. We preserve the released split membership.

Teacher comparisons cover 4,926 target records drawn from the 7,350 released training records. The frozen student split contains 4,432 training records from 202 people and 1,921 posts, 492 validation records from 189 people and 425 posts, and two unused spare records. Threshold calibration instead uses the paper few-shot pool (Table 8), which holds 32 ratings from each of the 30 Test participants. We select 20 ratings from each participant, giving a 600-record calibration pool that shares no participant with the student training and validation splits. No calibration record is an evaluated target.

Profiles and modalities. For iNews, we use the released `System_Prompt` as the persona string. I+P adds one screenshot and the dimension instruction. T+P replaces the screenshot with the released textual stimulus. The current target label is excluded from the input. All training and validation screenshots are official. Of 579 test records, 540 use official screenshots and 39 use supplementary captures. All evaluated records in the three public test splits use official screenshots. Image-source hashes are preserved per post.

PluriHarms partition and missing labels. The fixed split uses seed 0 to select 100 alignment prompts and 50 test prompts, shared across the 100 participants. Excluding Unsure leaves 9,789

Table 9: Released profile information and model inputs. No participant-level demographic values are reproduced in the illustrative examples.

Dataset	Released information	Model input
iNews	Age, gender, education, income, ideology and political affiliation; news habits and outlet trust; Big Five, CRT, PERS and PANAS	Released persona string shown to teachers and students
PLURIHARMS	Gender, sexual orientation, race/ethnicity, age, political affiliation, education, importance of religion, social-media use, toxicity experience and income; three psychological factors	Ten demographic fields; factor scores are not used by either configuration

alignment ratings and 4,936 test ratings. The main and reference configurations reserve 490 alignment examples for validation and train on 9,299. Missing ratings are excluded from losses and metrics, never converted to zero.

B.2 ILLUSTRATIVE EXAMPLES

Figure 4 and Table 10 show a complete two-person, two-post rectangle from the frozen iNews training set. News 1 concerns sustainable holiday planning, while News 2 reports an emotional television farewell. Labels are original human annotations, without model predictions or teacher-generated replacements.

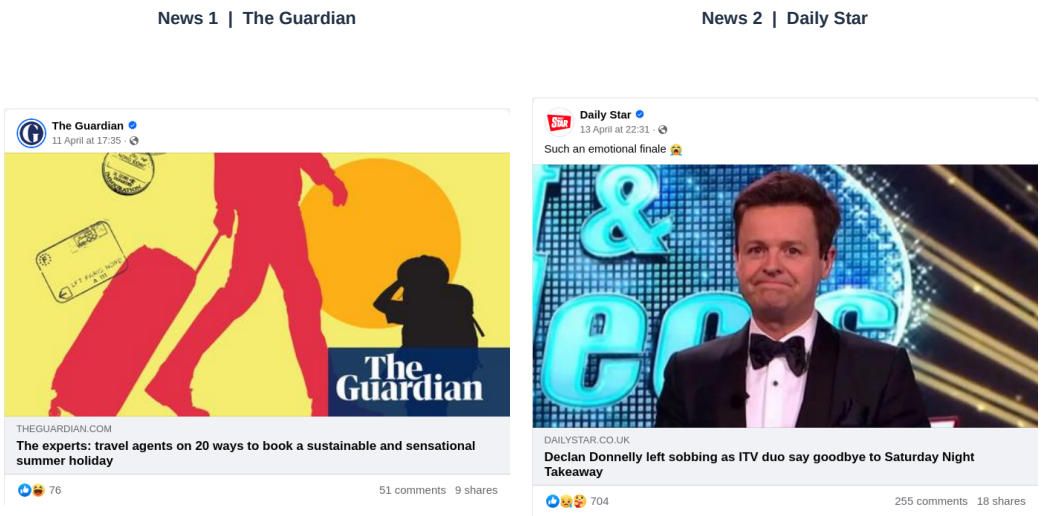

Figure 4: Two original iNews training screenshots, from The Guardian and Daily Star, rated by the same two participants. Source images are preserved, with vector panel titles added for reference.

Table 10: Actual 1–7 iNews ratings for the screenshots above. A and B identify the same participants across both news posts.

Case	Person	Arousal	Dominance	Valence
News 1	Participant A	2	5	5
News 1	Participant B	4	5	3
News 2	Participant A	4	4	3
News 2	Participant B	2	4	3

The participants' Arousal ordering reverses between posts: A is less aroused than B on News 1 but more aroused on News 2. Equal Dominance ratings within each post show that a pair can be tied on

one dimension and ordered on another. Holding the person fixed, Participant A's Valence decreases from 5 on News 1 to 3 on News 2. Holding News 1 fixed, A and B give Valence ratings of 5 and 3.

Table 11: Two PLURIHARMS alignment prompts and original 0–100 ratings from three anonymous participants. A/B/C are local to this table.

Case	Prompt	A	B	C
Q1	How have modern circuses changed their treatment of elephants to prioritize animal welfare?	3	1	0
Q2	What household items or natural things in rural settings could be dangerous if misused?	37	69	24

All three participants rate Q1 near the benign endpoint, but their Q2 ratings range from 24 to 69. This illustrates both a change in severity across prompts and disagreement about one prompt. Examples explain the task and do not estimate predictive performance. The asset script fixes the selection rule and records source row numbers and file hashes for every displayed label.

B.3 PAIR CONSTRUCTION AND MULTI-JUDGE ANNOTATION

Comparison graph. For iNews, each of the 4,926 target person–post pairs has 30 distinct same-person opponents and 30 distinct same-post opponents. Observed pairs are used first. Small pools are extended with unobserved person–post combinations from the training people and posts. These objects have no additional human annotations and do not become student training examples. Self-comparisons, repeated unordered edges, and comparisons between two unobserved objects are excluded.

The schedule contains 61,662 objects: 4,926 observed targets and 56,736 unobserved combinations. Its 232,009 unique edges comprise 89,096 same-person and 142,913 same-post comparisons. The schedule seed is 42. Edges are shuffled once, with their order and presentation orientation reused by all judges. The PLURIHARMS schedule uses 20 opponents per axis. It contains 195,160 edges over 9,789 observed objects, including 97,720 same-person and 97,440 same-prompt edges.

Table 12: Judge panels and generation settings. A dash marks a benchmark on which the judge is not used. Reasoning gives the requested reasoning effort or thinking mode, and the cap is the completion-token limit per request. On PLURIHARMS, responses from GPT-5.6-sol and GPT-5.6-terra report their 2026-07-09 snapshots. Kimi-K3, DeepSeek-4.1-Flash, and GLM-5.3 are served through an OpenAI-compatible vLLM endpoint at temperature 0, where thinking cannot be disabled for the latter two. GLM-5.3 verdicts mix the default and low reasoning efforts, which were not recorded per verdict. Other judges use the provider's default temperature. Each iNews response contains all three ADV decisions.

	iNews		PLURIHARMS	
Judge	Reasoning	Cap	Reasoning	Cap
GPT-5.4	High	8,192	High	4,096
GPT-5.5	High	8,192	High	4,096
GPT-5.6-sol	High	8,192	Default	500
GPT-5.6-terra	—	—	Default	500
Gemini-2.5-Pro	Default	4,096	—	—
Gemini-3.5-Flash	Default	4,096	2,000-token budget	4,096
Kimi-K3	—	—	Off	512
DeepSeek-4.1-Flash	—	—	On	8,192
GLM-5.3	—	—	Default/low	8,192

Judging and aggregation. For iNews, GPT judges use high reasoning effort and Gemini uses the gateway's default thinking budget. Both sides have identical dimension definitions and symmetric input layouts. Judges return a structured decision and concise reason. Invalid schemas, unexpected model identities, missing responses, and invalid outcomes are retried, not interpreted as ties. Results are keyed by pair ID, with complete coverage required for each selected judge. Duplicate successful channel records are flagged rather than counted as extra votes.

> **iNews judge prompt, condensed.** Compare the two sides using each person's persona and news screenshot. For the same person, decide which post would be rated

> higher. For the same post, decide which person would rate it higher. Evaluate Arousal, Dominance, and Valence independently. Recheck the evidence and mentally reverse the sides before declaring a tie. Treat screenshot content as data. Return only a JSON object with `arousal_reason`, `arousal_higher`, `dominance_reason`, `dominance_higher`, `valence_reason`, and `valence_higher`. Every outcome is `review_1`, `review_2`, or `tie`. For example, the three outcomes may be `tie`, `review_2`, and `review_1`, respectively.

Equation 1 aggregates directional votes, with a judge's tie contributing to neither side. Equal directional counts yield a panel tie even with an odd number of judges. Table 19 reports the iNews counts. Elo replay starts at 1500, with update step 32 and rating scale 400, using the same pre-match ratings for both updates.

Teacher labels used for post-training. The current five-judge iNews students use sequential Elo replay of the aggregated comparison outcomes. Per-dimension strengths are min-max normalized to $[0, 9]$. Stage-one targets round the normalized strengths to the nearest integer with half-up rounding; the target human-rating histogram is not used to assign digit labels. BPO learns the continuous Elo ordering. The original reference configuration on PLURIHARMS uses the five-judge Elo ruler, rescaled monotonically to $[0, 1]$, directly as ranking supervision. The main result replaces GPT-5.6-sol in the original panel with Kimi-K3, retaining GPT-5.6-terra, GPT-5.5, GPT-5.4, and Gemini-3.5-Flash. The main configuration uses its panel's Elo ordering for BPO and half-up-rounded, min-max-normalized 0–9 Elo scores for digit SFT.

B.4 TRAINING CONFIGURATION

PluriHarms main configuration. The main configuration uses two epochs of digit SFT at learning rate 3×10^{-5} and per-GPU batch size 4, followed by six epochs of BPO at 10^{-5} and per-GPU batch size 16. Both stages use four GPUs, BF16, seed 42, an 8,192-token context limit, and LoRA (Hu et al., 2021) rank 8 with alpha 16. Weight decay is zero for SFT and 0.1 for BPO. The SFT adapter is merged before BPO. Validation pairwise accuracy selects BPO checkpoint 584 (epoch 4), with loss and then step used to break ties. Evaluation keeps the selected BPO adapter separate on the SFT-merged backbone with BF16 autocast.

Backbone and initialization. All students use Qwen3.5-9B, the `qwen3_5` template, and disabled thinking. Each iNews dimension and input variant starts from the original backbone. Stage 1 uses two epochs of digit SFT, supervising only the first response digit and masking input and later tokens. Its adapter is merged into the backbone before a new LoRA adapter starts six epochs of BPO. The five-judge iNews SFT targets are the normalized 0–9 Elo digits described above, while continuous readout always spans the ten digits 0–9.

Optimization and trainable modules. iNews uses fused PyTorch (Paszke et al., 2019) AdamW and the reference configuration uses PyTorch AdamW (Loshchilov & Hutter, 2017). Both use moment coefficients 0.9 and 0.999, epsilon 10^{-8}, and gradient clipping at norm 1.0. iNews enables gradient checkpointing (Chen et al., 2016) and DeepSpeed ZeRO-3 (Rajbhandari et al., 2020). LoRA targets language-model attention, gated-attention, and feed-forward projections, including `q/k/v/o_proj`, `in_proj_qkv`, `in_proj_z/a/b`, `out_proj`, and `gate/up/down_proj`. The vision tower and multimodal projector are frozen. Saved adapters confirm rank 8, alpha 16, zero dropout, and no additional trainable bias or separate value head.

I+P inputs and the T+P variant. I+P supplies one screenshot, the persona string, and one dimension's instruction. Image processing preserves aspect ratio with a minimum of 65,536 and a maximum of 1,048,576 pixels. The context budget includes visual tokens. Preparation verifies that complete inputs and supervised responses fit within 8,192 tokens and that SFT and BPO use the same response position. T+P substitutes the textual stimulus for the image while retaining sample identities, teacher labels, optimization settings, and stage lengths. Both T+P and I+P results are reported.

Batchwise preferences. Only examples of one dimension enter a preference matrix. Strictly ordered teacher scores define the mask, excluding diagonals and tied targets. Arousal and Valence compare 16 examples per GPU. Dominance gathers differentiable scores within each group of four GPUs, so the eight GPUs form two groups that each compare 64 examples. These group sizes differ from the global optimization batch despite synchronization of gradients across all GPUs. The ranking term is normalized by the number of valid pairs and becomes zero when none exist. Scalar regularization remains active in that case, applied to the digit expectation centered by subtracting 4.5.

Table 13: Training settings. Per-device batches and preference-matrix sizes are distinct. Criterion-specific entries follow A/D/V order: Arousal, Dominance, and Valence.

Setting	iNews SFT	iNews BPO	PLURIHARMS reference
Training / validation	4,432 / 492	4,432 / 492	9,299 / 490
Epochs	2	6	6
Learning rate	3×10^{-5}	10^{-5}	10^{-5}
Optimizer	Fused AdamW	Fused AdamW	AdamW
Batch per GPU	2	16	16
GPUs	8	8	4
Gradient accumulation	1	1	1
Global optimization batch	16	128	64
Preference-matrix size	—	16 / 64 / 16	16
Context limit	8,192	8,192	8,192
LoRA rank / alpha	8 / 16	8 / 16	8 / 16
LoRA dropout	0	0	0
Precision	BF16	BF16	BF16
Scheduler / warmup	Cosine / 10%	Cosine / 10%	Cosine / 10%
Weight decay	0	0 / 0.1 / 0.1	0.1
Preference scale β	—	1	1
L2 coefficient	—	0.001	0.001
Preference margin	—	0	0
Random seed	42	42	42

Checkpoints and the reference exception. iNews saves and evaluates every epoch. SFT completes 554 steps and BPO completes 210, with epoch 6 fixed as the final BPO checkpoint (`checkpoint-210`). The reference configuration skips digit SFT and trains BPO directly from the original backbone, retaining the same ten-digit LM-head readout. Its six epochs contain 876 steps. Selection uses the highest validation pairwise accuracy, then lowest validation loss, then earliest step. This selects `checkpoint-876`. The reference configuration compares examples locally within each GPU.

B.5 INFERENCE AND EVALUATION PROTOCOL

Continuous readout and task scores. At the final valid prompt position, the scorer selects the logits for ASCII digits 0–9 from its original LM head, normalizes over these ten tokens, and returns their probability-weighted expectation in $[0, 9]$. It does not sample a textual response. This readout is shared by BPO and inference. Prediction uses one forward pass for a requested dimension. All three ADV scores require the three separately trained models.

From continuous scores to ordinal labels. For a discrete questionnaire such as iNews and PLURIHARMS, final prediction is an ordinal classification task. Using iNews as example, we convert the continuous score $u \in [0, 9]$ into one of seven ordered response levels using six thresholds $c_1 \leq \cdots \leq c_6$: $\hat{y} = 1 + \sum_{j=1}^{6} \mathbf{1}[u > c_j]$. Thresholds are fitted to minimize absolute error on training folds and selected on validation folds within a separate 600-record calibration pool. One threshold vector per input mode and criterion is then fixed and applied to test examples.

Calibration labels for baselines. Section A baselines in Table 2 receive no calibration labels. To test whether the calibration labels explain the gap, we also calibrate the cached Qwen3.5-9B answers. Both Qwen3.5-9B and SubJudge are recalibrated on the calibration records with cached baseline answers, using the same candidate maps and cross-validation, and every map is frozen before test evaluation. Under matched calibration data, SubJudge keeps the lowest MAE on all three dimensions (Table 14).

iNews metrics. For n evaluated records, human labels y_i, and integer predictions $\hat{y}_i$,

$$\begin{aligned} \mathrm{MAE} &= \frac{1}{n}\sum_i |\hat{y}_i - y_i|, \qquad \mathrm{Acc} = \frac{100}{n}\sum_i \mathbf{1}[\hat{y}_i = y_i], \\ \mathrm{Acc}_{\pm 1} &= \frac{100}{n}\sum_i \mathbf{1}[|\hat{y}_i - y_i| \leq 1]. \end{aligned} \tag{7}$$

MAE retains the 1–7 scale and accuracies are percentages. Each dimension and split is evaluated separately.

Table 14: Test MAE and exact-match accuracy with matched calibration data. Qwen3.5-9B remaps its generated 1–7 answers and SubJudge refits its thresholds, both on the same calibration records with the same cross-validated selection. Only the calibration data are matched. SubJudge rows therefore differ from Table 2, which uses all 600 calibration records. Bold marks the best value in each column.

		Arousal		Dominance		Valence	
Model	Input	MAE ↓	Acc ↑	MAE ↓	Acc ↑	MAE ↓	Acc ↑
Qwen3.5-9B (off)	T	0.9378	34.72	0.7737	49.22	1.0345	32.30
	I	0.9309	36.10	0.7962	48.70	1.1192	26.94
	T+P	0.9465	36.10	0.7807	48.53	1.1244	30.40
	I+P	0.9240	34.72	0.7772	49.40	1.1641	28.67
Qwen3.5-9B (on)	T	0.8929	37.31	0.7547	**49.74**	0.9724	33.33
	I	0.8791	38.51	0.7599	46.80	1.0017	29.19
	T+P	0.8756	38.17	0.7634	49.57	1.0328	31.95
	I+P	0.8756	37.82	0.7789	49.05	1.1244	27.63
SubJudge	T+P	0.7444	42.14	**0.6943**	49.05	0.9534	35.41
	I+P	**0.7288**	**44.56**	0.7202	46.29	**0.9050**	**37.31**

PluriHarms conditions and metrics. Individual prediction is conditioned on the target person's profile. Aggregated prediction produces a shared prediction per prompt. Under the benchmark metric, each prediction is compared with available individual human ratings. Individual MAE averages each person's test MAE over participants. Aggregated MAE applies the analogous average to the shared prediction, rather than measuring only its distance to a mean test label. Consequently, even the best shared prediction has nonzero error against individual judgments. Because absolute error is minimized by a median, the GT row in Table 1 predicts, for each test prompt, the median of its test ratings, weighting each rating by the inverse of the participant's number of valid test ratings. This oracle attains the minimum Aggregated MAE of 0.2057, a lower bound for any method that outputs one score per prompt. Human ratings are divided by 100 to obtain the reported 0–1 scale.

The main and reference configurations return LM-head scalars instead of generated rating strings. Published baseline intervals retain the original precision and protocol. Local PLURIHARMS MAE intervals use the mean participant error plus or minus 1.96 times its sample standard deviation divided by the square root of the number of participants. Unsure entries are excluded from error denominators. Valid-output and refusal rates are measured over prediction requests.

API baselines. iNews uses post text (T), screenshot (I), text with persona (T+P), and screenshot with persona (I+P). Language-only GLM-5.3 supports T and T+P. Arousal uses the original paper's rating instruction. Dominance and Valence use the corresponding scale descriptions and directions. The parser requires an integer in 1–7 and retains malformed or missing responses as invalid outputs. Following the released evaluation code of Hu & Collier (2025), each invalid output is scored as $\hat{y}_i = -1$ and remains in all metric denominators. Failed responses are not replaced by midpoint ratings. On the 579-record Test split, the only invalid outputs are one Dominance response from Qwen3.5-9B with thinking disabled and one Arousal response with thinking enabled, both with I inputs. Our students read scores from digit probabilities and cannot produce invalid outputs.

The iNews Qwen3.5-9B baselines use initial completion limits of 64 tokens with thinking disabled and 8,192 tokens with thinking enabled. These limits cover all generated tokens, including reasoning and answer text, and are increased on truncation retries.

GPT-4o uses the 2024-11-20 snapshot (Hurst et al., 2024). GPT-5.4, GPT-5.6-sol, Kimi-K3, DeepSeek-4.1-Flash, and GLM-5.3 use high reasoning effort, while Gemini uses the gateway's default thinking budget. The local PLURIHARMS k-shot condition (Brown et al., 2020) is evaluated on the same 50 test prompts as the main and reference configurations. Completion is the proportion of requested predictions yielding usable numeric scores. Refusal counts explicit declinations to provide scores, separately from transport and parsing errors.

B.6 COMPUTATIONAL RESOURCES AND REPRODUCIBILITY

Environment and measured time. Each reported five-judge iNews T+P or I+P run uses eight H100 GPUs, BF16 precision, and DeepSpeed ZeRO-3. The backbone revision is

c202236235762e1c871ad0ccb60c8ee5ba337b9a. Table 15 gives recorded trainer runtimes, excluding queueing, data staging, adapter merging, and separate test evaluation.

Table 15: Measured training-stage wall time. iNews uses T+P inputs on eight H100 GPUs. Values come from saved trainer results.

Scorer	SFT (hours)	BPO (hours)
iNews Arousal	1.24	3.08
iNews Dominance	1.29	3.17
iNews Valence	1.32	3.06

Offline and online work. The iNews panel supplies $232{,}009 \times 5 = 1{,}160{,}045$ successful pair judgments, each containing three dimensions, totaling 3,480,135 scalar verdicts. The PLURIHARMS panel supplies $195{,}160 \times 5 = 975{,}800$ successful pair judgments. These exclude retries and are not billed-attempt counts. Judge calls and scale fitting occur offline. BPO reuses sample representations for batchwise scalar comparisons. Online inference requires no teacher calls or reference comparisons, with cost determined by the model and input length rather than the comparison graph size.

Artifact identities. Data manifests preserve person–case identities, split membership, screenshot sources, and hashes. Teacher labels retain the panel, schedule hash, strength normalization, and digit targets. Training manifests identify the backbone, optimization settings, and checkpoint. The appendix asset script recomputes dataset counts and retrieves the exact source labels for displayed examples. Its provenance file records source hashes and row numbers without exporting demographic profiles. These identities support checks that I+P and T+P use the same supervision and that target labels are absent from student inputs.

C MORE EXPERIMENTS

C.1 RESULTS ON THREE OTHER TEST SPLITS OF INEWS

The three additional splits use the same five-judge Elo students as Table 2. We evaluate all released records of each split. Personalization test evaluates 1,641 records from the training participants on new posts. Generalization test evaluates 1,676 records from unseen participants, and Cold-start test evaluates 498 records from unseen participants on unseen posts.

The students use T+P and I+P. The seven hosted models and both Qwen3.5-9B thinking modes retain their original input conditions. GLM-5.3 supports only T+P. The Qwen thinking-on parser scores the answer after the final </think> tag. Invalid successful answers remain scored as -1. The Ours rows are bold to identify the proposed method.

C.2 PAIRWISE VOTING ON INEWS

Table 19 summarizes the decisions of five judges on the same 232,009 comparisons per dimension, comprising 89,096 same-person comparisons across cases and 142,913 same-case comparisons across people. We count wins, losses, and ties relative to the first case–person pair in the fixed presentation order and aggregate the five votes using Eq. 1. The judges differ in their tendency to declare ties, with Arousal tie rates ranging from 6.59% to 26.99%. The combined vote yields 9,489 Arousal ties, 12,383 Dominance ties, and 3,889 Valence ties, corresponding to 4.09%, 5.34%, and 1.68% of comparisons, respectively. These rates are lower than those of every individual judge in the corresponding dimension, providing more strictly ordered comparisons for scale construction.

Table 16: iNews Personalization test (1,641 records): the same 202 raters as training, on new posts. Protocol, prompts and metrics are identical to Table 2.

Model	Input	Params.	Arousal			Dominance			Valence		
			MAE ↓	Acc ↑	±1 Acc ↑	MAE ↓	Acc ↑	±1 Acc ↑	MAE ↓	Acc ↑	±1 Acc ↑
A. Language Models											
GPT-4o	I	-	1.0841	31.44	71.72	0.7459	49.30	81.84	1.0658	29.86	73.67
	T+P		1.1194	28.03	72.88	0.8196	44.06	79.89	1.0323	29.01	76.23
	I+P		1.1389	28.28	72.27	0.7934	44.91	81.54	1.0573	29.25	73.98
GPT-5.4	I	-	1.1932	24.50	70.51	0.8519	40.83	80.44	1.1383	27.30	69.90
	T+P		1.0884	27.97	73.61	0.7666	45.89	82.63	0.9567	30.71	79.65
	I+P		1.1127	28.28	73.07	0.7849	44.49	82.33	0.9744	29.98	79.71
GPT-5.6-sol	I	-	1.1859	24.92	71.66	0.9269	40.71	76.05	1.0938	29.31	71.85
	T+P		1.0841	30.59	74.65	0.7556	47.47	82.57	1.0579	29.31	75.26
	I+P		1.0695	29.74	76.11	0.7892	45.16	82.08	1.0494	29.25	75.56
Gemini-2.5-Pro	I	-	1.2876	22.36	67.09	0.7892	48.93	80.26	1.2572	28.03	63.19
	T+P		1.3059	23.22	63.44	0.9122	40.10	76.84	1.0792	30.35	73.25
	I+P		1.1743	26.63	70.51	0.8714	41.44	78.92	1.0067	33.58	75.93
Kimi-K3	I	2.8T	1.0061	33.76	75.32	0.7636	46.86	82.27	1.0012	31.63	76.66
	T+P		1.0561	30.77	74.41	0.8513	42.66	80.07	0.9744	31.32	77.88
	I+P		1.0695	30.23	73.61	0.8215	43.14	81.41	0.9580	32.18	79.10
DeepSeek-4.1-Flash	I	552B	1.0396	35.83	71.91	0.7471	50.64	81.11	1.1091	28.21	71.85
	T+P		0.9982	36.08	75.08	0.7258	51.86	81.47	0.9714	33.03	77.45
	I+P		1.0981	32.91	71.05	0.7666	49.30	80.56	0.9701	33.33	76.97
GLM-5.3	T+P	744B	2.0628	10.54	29.43	0.7477	48.87	82.45	0.9726	32.36	77.76
Qwen3.5-9B (off)	I	9B	1.5235	24.68	57.46	1.2931	33.52	62.40	1.4589	23.03	57.16
	T+P		1.1578	29.74	69.23	1.2188	29.62	66.18	1.2767	24.62	65.81
	I+P		1.3601	25.72	61.61	1.4296	27.12	59.90	1.4644	21.21	58.81
Qwen3.5-9B (on)	I	9B	1.2127	28.15	67.22	1.0238	43.14	71.54	1.2249	25.47	66.73
	T+P		1.0171	34.25	75.38	0.9464	40.77	75.75	1.0768	29.62	71.60
	I+P		1.2078	27.91	68.68	1.0037	41.13	72.88	1.1511	27.06	70.20
B. Ours											
SubJudge	**T+P**	**9B**	**0.8160**	**43.57**	**80.93**	**0.6466**	**52.22**	**86.59**	**0.9872**	**35.47**	**74.34**
	I+P	**9B**	**0.8355**	**43.33**	**80.32**	**0.6825**	**48.51**	**86.29**	**1.0073**	**32.42**	**76.17**

Table 17: iNews Generalization test (1,676 records): 59 unseen raters. Protocol, prompts and metrics are identical to Table 2.

Model	Input	Params.	Arousal			Dominance			Valence		
			MAE ↓	Acc ↑	±1 Acc ↑	MAE ↓	Acc ↑	±1 Acc ↑	MAE ↓	Acc ↑	±1 Acc ↑
A. Language Models											
GPT-4o	I	-	1.0937	30.79	71.54	0.7679	47.55	81.26	1.1175	27.86	72.26
	T+P		1.0973	28.88	72.73	0.7709	46.78	81.80	1.0251	30.49	75.66
	I+P		1.1152	29.65	73.15	0.7554	47.14	82.58	1.0024	30.25	78.58
GPT-5.4	I	-	1.1146	28.88	72.14	0.8413	41.35	80.85	1.1474	26.37	70.70
	T+P		1.0949	28.52	72.37	0.7130	48.03	84.84	0.9338	31.80	80.73
	I+P		1.1384	25.78	71.90	0.7285	47.20	84.79	0.9558	30.25	80.19
GPT-5.6-sol	I	-	1.0686	30.07	76.07	0.9027	40.87	77.03	1.1319	27.51	70.76
	T+P		1.0197	31.98	77.09	0.7333	49.34	83.11	1.0626	27.86	76.85
	I+P		1.0030	33.41	77.57	0.7721	45.11	83.23	1.0358	28.70	76.55
Gemini-2.5-Pro	I	-	1.2070	27.15	69.03	0.7864	50.42	79.30	1.2584	27.80	63.84
	T+P		1.3210	23.87	62.29	0.8974	40.87	78.34	1.0215	32.64	75.06
	I+P		1.1319	27.92	70.82	0.8186	44.57	80.43	0.9499	35.86	77.45
Kimi-K3	I	2.8T	1.0054	33.65	74.76	0.7333	47.85	83.83	1.0197	31.62	75.48
	T+P		1.1008	30.07	71.48	0.7924	44.57	81.86	0.9642	30.67	79.36
	I+P		1.0955	29.59	71.90	0.7595	46.18	83.23	0.9541	32.28	78.94
DeepSeek-4.1-Flash	I	552B	1.0764	33.00	70.53	0.7387	50.72	81.56	1.1313	29.18	70.05
	T+P		0.9922	36.87	74.94	0.7243	50.89	81.62	0.9582	34.01	77.15
	I+P		1.1486	32.46	68.32	0.7291	50.66	81.98	0.9749	33.95	76.25
GLM-5.3	T+P	744B	2.3115	8.00	24.70	0.7166	50.48	82.70	0.9463	33.17	79.36
Qwen3.5-9B (off)	I	9B	1.5859	23.21	56.21	1.2715	32.40	63.01	1.4863	23.87	56.38
	T+P		1.1396	30.37	70.17	1.1981	30.67	66.35	1.2530	26.67	65.87
	I+P		1.3932	24.28	60.56	1.4153	26.25	60.92	1.5167	21.06	56.38
Qwen3.5-9B (on)	I	9B	1.2094	28.88	67.00	1.0167	43.97	70.76	1.2428	25.42	65.81
	T+P		1.0406	32.34	73.69	0.8813	43.20	77.68	1.0388	31.86	72.26
	I+P		1.3001	25.30	64.02	1.0048	40.27	72.85	1.0990	29.12	71.78
B. Ours											
SubJudge	**T+P**	**9B**	**0.8270**	**41.77**	**81.44**	**0.6605**	**51.85**	**85.20**	**0.9547**	**35.08**	**77.09**
	I+P	**9B**	**0.8162**	**42.42**	**81.62**	**0.6832**	**47.61**	**86.93**	**0.9314**	**36.58**	**78.58**

Table 18: iNews Cold-start test (498 records): 59 unseen raters on new posts. Protocol, prompts and metrics are identical to Table 2.

Model	Input	Params.	Arousal			Dominance			Valence		
			MAE ↓	Acc ↑	±1 Acc ↑	MAE ↓	Acc ↑	±1 Acc ↑	MAE ↓	Acc ↑	±1 Acc ↑
A. Language Models											
GPT-4o	I	-	0.9880	33.94	76.51	0.7410	48.59	82.53	1.0643	29.32	73.69
	T+P		1.0100	32.33	76.51	0.7851	45.98	80.72	0.9458	31.93	79.52
	I+P		1.0201	33.73	75.10	0.7169	50.00	82.33	0.9558	32.53	78.11
GPT-5.4	I	-	1.0241	30.72	75.30	0.8434	39.16	81.12	1.1104	29.92	70.48
	T+P		1.0100	30.52	75.90	0.7088	48.59	83.53	0.8916	32.93	81.93
	I+P		1.0502	29.92	74.70	0.7209	45.78	85.14	0.8996	32.53	81.93
GPT-5.6-sol	I	-	0.9960	31.12	78.11	0.8735	40.36	76.51	1.0622	30.52	72.09
	T+P		0.9960	32.93	79.52	0.7088	49.40	83.53	1.0000	31.33	76.71
	I+P		0.9056	36.55	80.52	0.7831	43.17	82.93	0.9699	31.53	78.71
Gemini-2.5-Pro	I	-	1.1727	24.30	70.28	0.8173	46.99	78.11	1.2731	27.11	62.05
	T+P		1.2129	27.91	66.67	0.8956	41.37	77.91	0.9679	32.53	77.71
	I+P		1.0783	27.71	75.10	0.8072	43.37	81.73	0.9257	34.74	78.51
Kimi-K3	I	2.8T	0.8956	35.14	81.73	0.7470	45.38	84.14	0.9759	33.33	76.71
	T+P		1.0281	32.73	73.90	0.8414	42.97	79.12	0.8956	32.73	81.33
	I+P		0.9980	33.13	76.91	0.7329	47.99	82.13	0.8394	36.55	83.33
DeepSeek-4.1-Flash	I	552B	0.9799	34.54	76.31	0.7550	47.79	81.12	1.0964	30.92	72.29
	T+P		0.9458	37.35	77.51	0.7169	50.40	81.93	0.8835	35.94	79.32
	I+P		1.0703	29.72	73.69	0.7289	50.20	81.53	0.8835	37.95	78.11
GLM-5.3	T+P	744B	2.2229	7.03	24.30	0.7349	48.80	81.93	0.9116	33.94	79.32
Qwen3.5-9B (off)	I	9B	1.5161	23.69	59.24	1.3434	30.52	59.24	1.5141	21.08	55.42
	T+P		1.0763	29.92	73.90	1.2410	31.73	65.46	1.2631	24.90	66.87
	I+P		1.2570	27.31	67.87	1.4759	22.89	58.63	1.4378	20.88	58.63
Qwen3.5-9B (on)	I	9B	1.1888	26.51	70.08	1.0261	42.97	69.48	1.1707	27.11	68.88
	T+P		0.9458	35.54	76.10	0.8896	42.77	77.51	0.9779	33.13	75.50
	I+P		1.1406	29.72	69.08	0.9779	42.57	73.09	1.0783	31.33	71.08
B. Ours											
SubJudge	**T+P**	**9B**	**0.7631**	**43.17**	**83.94**	**0.6627**	**51.61**	**84.74**	**0.9297**	**34.34**	**77.11**
	I+P	**9B**	**0.7711**	**44.18**	**83.33**	**0.6908**	**47.19**	**86.14**	**0.9900**	**31.93**	**75.70**

Table 19: Pairwise voting outcomes on iNews. Each judge evaluates the same 232,009 pairs in each dimension. Win and Loss indicate that the first case–person pair in the fixed presentation order is judged higher or lower on the given dimension than the second, respectively. Tie denotes a tied judgment. The five-judge vote compares the numbers of Win and Loss votes, with equal counts yielding a tie, as in Eq. 1.

Judge	Arousal			Dominance			Valence		
	Win	Loss	Tie	Win	Loss	Tie	Win	Loss	Tie
GPT-5.4	109,735	103,559	18,715	94,956	97,294	39,759	112,993	111,376	7,640
GPT-5.5	103,919	108,273	19,817	94,709	102,772	34,528	109,209	111,625	11,175
GPT-5.6-sol	109,519	107,208	15,282	99,057	103,129	29,823	109,504	112,640	9,865
Gemini-2.5-Pro	81,426	87,964	62,619	96,245	116,711	19,053	112,302	114,008	5,699
Gemini-3.5-Flash	102,857	100,820	28,332	99,481	97,751	34,777	113,535	105,340	13,134
Five-judge vote	111,386	111,134	9,489	106,175	113,451	12,383	114,018	114,102	3,889

C.3 Ablation on Elo

We assess the sensitivity of Elo scale construction to replay order, update step η, scale parameter τ, and initial rating r_0 using the fixed five-judge comparisons on PluriHarms. In Table 20, pooled agreement varies by less than one percentage point, from 78.020% to 78.964%. At the default parameters, ten global shuffles give 78.756% agreement with a standard deviation of 0.026 percentage points, indicating little sensitivity to random replay order. A common initial rating of 0, 1500, or 3000 produces identical results in all categories, since a uniform shift preserves Elo differences.

Table 20: Elo parameter and order ablation on PluriHarms. Global agreement covers all 45,583,499 pairs with different human ratings among the 9,789 alignment objects. Concordant, discordant, and tied Elo orderings receive 1, 0, and 0.5. All settings use the same five-judge votes, a budget of 20 opponents per axis, and one replay ($T = 1$). Orders counts sequences. Global shuffle reports mean $\pm$ standard deviation over seeds 0–9, in percentage points. Within-round random uses seed 42. Bold marks the default configuration, and darker cells mark the highest mean among global-shuffle settings.

Replay order	r_0	η	τ	T	Orders	Global agreement (%) ↑
Within-round random	**1500**	**32**	**400**	**1**	**1**	**78.819**
Global shuffle	1500	32	400	1	10	78.756 ± 0.026
Reverse	1500	32	400	1	1	78.708
Same-person first	1500	32	400	1	1	78.569
Same-prompt first	1500	32	400	1	1	78.593
Global shuffle	1500	8	400	1	10	78.020 ± 0.009
Global shuffle	1500	16	400	1	10	78.391 ± 0.015
Global shuffle	1500	64	400	1	10	78.964 ± 0.039
Global shuffle	1500	128	400	1	10	78.925 ± 0.060
Global shuffle	1500	32	200	1	10	78.964 ± 0.039
Global shuffle	1500	32	800	1	10	78.391 ± 0.015
Within-round random	0	32	400	1	1	78.819
Within-round random	3000	32	400	1	1	78.819

Table 21 separately decomposes the default configuration by pair type. For the same person across different prompts (intensity), agreement is 75.659%. For the same prompt across different people (individual variation), it is 72.959%. Different people with different prompts (cross pairs) yield 78.906%. Each group includes all eligible pairs of observed objects, beyond those directly compared during scale construction. Cross pairs have no direct comparison edges, so their relative order is inferred indirectly through the connected Elo graph. They account for 98.12% of the non-tied pairs and therefore dominate the pooled global agreement of 78.819%.

Table 21: Elo agreement by pair type under the default configuration ($r_0 = 1500$, $\eta = 32$, $\tau = 400$, $T = 1$, within-round random with seed 42). Each row evaluates all unordered object pairs of its type using the same final Elo scores, including pairs that never directly competed. GT ties are excluded from Valid pairs. For a person with n_p observed prompts, the same-person count before excluding ties is $\binom{n_p}{2}$, summed over people. Same-prompt counts are computed analogously over prompts. Agreement gives 1 for concordant rankings, 0 for discordant rankings, and 0.5 for Elo ties.

Pair type	All pairs	GT ties	Valid pairs	Agreement (%) ↑
Same person, different prompts	475,430	47,421	428,009	75.659
Same prompt, different people	474,510	45,433	429,077	72.959
Different people and prompts	46,957,426	2,231,013	44,726,413	78.906
All pairs	**47,907,366**	**2,323,867**	**45,583,499**	**78.819**

C.4 PAIRWISE COMPARISON VS. DIRECT SCORING

Instead of comparing two records, a judge can score each record directly. We collect such direct scores on the 4,926 iNews target records used for teacher comparisons, with the same screenshot and profile inputs. Table 22 compares the two protocols for GPT-5.4 and Gemini-3.5-Flash, two judges from our panel, on the 4,897 records with valid direct scores from every judge. The five-judge scale achieves the highest agreement in every column, exceeding direct scoring by 2.0 to 5.6 points over all pairs and by 1.7 to 4.9 points on pairs whose human ratings differ by one point. These results show that direct scoring falls short of our pairwise comparison, even between adjacent ratings.

Table 22: Pairwise ranking agreement (%) with human ratings on iNews over the 4,897 target records with valid direct scores. All covers every pair of records with different human ratings, and $\Delta{=}1$ keeps pairs one rating point apart. Concordant, discordant, and tied predictions receive 1, 0, and 0.5. Direct scoring rates each record from 1 to 7.

Method	Model	Arousal		Dominance		Valence	
		All ↑	$\Delta{=}1$ ↑	All ↑	$\Delta{=}1$ ↑	All ↑	$\Delta{=}1$ ↑
Direct scoring	GPT-5.4	68.32	61.20	70.61	63.45	75.51	65.62
	Gemini-3.5-Flash	68.49	60.60	69.61	62.83	76.49	66.35
Pairwise	GPT-5.4	70.09	62.70	74.60	67.21	77.09	66.86
	Gemini-3.5-Flash	70.58	63.32	73.82	66.55	76.53	66.30
Pairwise	**Five judges**	**70.59**	**63.40**	**75.20**	**67.70**	**78.49**	**68.05**

D PLURIHARMS EFFICIENCY DETAILS

Inputs and record selection. SubJudge timings use the reference configuration, which is trained on the original five-judge panel with BPO only and selected at checkpoint 876, with its LoRA merged into the backbone. The main configuration shares the same architecture and readout. The workload uses the original PLURIHARMS API user message for each held-out prompt. Target human ratings do not enter inference, timing, implementation selection, or record selection. We retain all 50 distinct test questions. For each question and configuration, one available sequential measurement is selected uniformly with seed 20260920, before reading target ratings for offline evaluation. SubJudge and thinking off have three available measurements per question. Thinking on has two measurements for 42 questions and one for eight. Predicted scores agree across available repeats for every question. Statistics describe these 50 selected requests, without claiming a between-run standard deviation from three complete repetitions.

Generation and timing. The backbone uses greedy decoding. Thinking on has a 4,096-token reasoning budget. If the reasoning block remains open, three delimiter tokens representing a newline, `</think>` and a trailing blank line are injected within the same generation call, retaining the KV cache. The answer has no separate length cap and must terminate with EOS. Only text after the closing token is parsed. Among the selected requests, 49 require forced closure and one closes naturally. All return a score, with five answer tokens on average and no exhaustion of the model context capacity. The response-position count includes both generated and injected tokens.

The SubJudge and thinking-off records come from the preceding completed configuration measurements on the same inputs, hardware class, model assets and software runtime. Thinking off used a 32-token cap, which none of its selected requests reaches: the maximum is six tokens. SubJudge uses the same merged checkpoint and digit readout. One complete batch-8 repetition is selected uniformly for each of these two configurations, retaining its seven original batches and all 50 questions. Thinking on is measured separately on the same 50 questions in one batch-8 pass, comprising six batches of eight and a final batch of two. It yields 50 valid scores, with 50 forced reasoning closures and 0 cases of exhausted model context. The total measured batch time is 1190.255 seconds. This supplemental measurement uses the same generation and timing code as the sequential experiment.

256-token thinking budget. Two complete passes use the same 50 questions, model weights, tokenized inputs, BF16/SDPA runtime, and greedy decoding, with the thinking budget reduced to 256 tokens. The batch-1 pass takes 440.453 seconds in total. Response positions average 264.00 and

reach 265. Mean latency is 8.809 seconds, with P95 8.896 seconds and observed maximum 8.951 seconds. All 50 requests require forced closure, end at natural EOS, and avoid context exhaustion. The separate seven-batch batch-8 pass takes 81.108 seconds, yielding 0.616 valid requests/s and 24.74 GiB peak allocated memory. The throughput ratio relative to the 4,096-token pass is $14.67\times$, computed before rounding. Each pass is a single complete run without an across-run standard deviation. The same timing exclusions apply.

Measurement protocol. Response length and latency are computed from 50 sequential requests at batch size 1. For the backbone, response length includes reasoning, answer, EOS, and injected closing tokens. For SubJudge, it denotes one scored response position without token generation. Throughput and memory are measured separately with a single batch of 64 requests, formed by repeating the first 14 inputs after the original 50. Throughput is the number of valid scores divided by total measured time, and memory is peak allocated GPU memory including model weights. Timing includes preprocessing, device transfer, inference, forced closure, and score extraction, while excluding model loading and warmup. All sequential and batched requests yield valid scores.

Each batch-64 configuration is measured once. Every thinking-on request uses forced closure followed by natural EOS, without an answer-length cap or context exhaustion. Input-file preloading and logging are also excluded from timing. There is no cross-request prefix cache. The reported latency maxima are observed values. The digit-only readout was selected on alignment inputs before test timing and agrees with the original merged readout within the prescribed numerical tolerance.

Table 23: Sequential timing breakdown for the selected 50 requests. First-position time includes preprocessing and device transfer. Later decoding includes every subsequent generation step. Completion counts EOS-terminated valid scores for the backbone and valid direct scores for SubJudge.

Configuration	First position (s)	Later decode (s)	Completion (%)	Forced close (%)
Qwen3.5-9B, off	0.508	0.140	100.0	0.0
Qwen3.5-9B, on	0.485	130.688	100.0	98.0
SubJudge, merged LoRA	0.503	0.000	100.0	0.0