WHEN DO LLM JUDGE PANELS OUTPERFORM A SINGLE LARGE JUDGE? THE DIVERSITY-BIAS TRADEOFF

Anonymous authors
Paper under double-blind review

ABSTRACT

LLM evaluation pipelines increasingly face a practical choice: use one strong judge or aggregate a panel of diverse judges. The usual argument for panels is variance reduction through less-correlated errors, but this view ignores an opposing effect: provider diversity can also increase bias heterogeneity. We formalize this *diversity-bias tradeoff* with a two-term decomposition of panel gain and derive a phase boundary characterizing when aggregation should outperform a single judge. Across four independently composed cross-provider panels spanning seven AI providers and four NLG benchmarks (21 dimensions, 2,635 items), provider diversity reduces inter-judge error correlation on every dimension tested ($\bar{\rho} = 0.493$–0.726 versus 0.925 for a same-provider baseline), yet lower correlation does not translate monotonically into higher evaluation accuracy because bias heterogeneity can offset the variance benefit. The decomposition predicts the direction of panel gain on 94% of dimensions satisfying its same-sign-bias assumption (81% overall). An equal-budget comparison finds a diverse panel approximately equivalent to a comparably priced single judge (10 versus 11 wins across 21 dimensions), showing that apparent panel advantages can reflect judge quality rather than diversity alone. Item-level analysis provides additional directional support for the variance-reduction mechanism (Spearman $r = -0.10$, $p < 0.0001$, $N = 4{,}200$). Finally, we translate the theory into a validation-based deployment rule that selects between panel and single-judge evaluation at the dimension level. These results explain when judge diversity helps, when heterogeneous bias cancels its benefit, and why panel deployment should be selective rather than automatic.

1 INTRODUCTION

LLM-as-judge has become a practical foundation for scalable language-model evaluation (Zheng et al., 2023; Liu et al., 2023; Chiang et al., 2024). As these evaluations grow in volume and cost, practitioners face a recurring design choice: spend the budget on one strong judge, or aggregate a panel of smaller, diverse judges? The case for panels appears straightforward: if judges make imperfect but sufficiently independent errors, aggregation should cancel noise, and Verga et al. (2024) show that panels of smaller models can match or outperform a single large judge at lower cost. But recent work exposes a complication: LLM judges are often far from independent. Kohli (2026) find that nine judges may provide only approximately two independent votes of information; Kim et al. (2025) show strong same-provider error correlation; and Jiang et al. (2025) document broad output homogenization across frontier models.

These findings motivate provider diversity as an apparent remedy: combine judges from different model families to reduce correlated error. Yet a basic question remains: *when does that diversity actually improve the final judgment?* Lower correlation helps only if it is not purchased at the cost of larger systematic disagreement with the target, since judges from different providers may make less-correlated errors precisely because they encode different scoring tendencies and calibration biases. A panel wins only when the variance benefit exceeds this bias cost.

We formalize this tension as the **diversity-bias tradeoff**. For bounded multi-attribute NLG evaluation, we derive a two-term decomposition of panel gain that separates the benefit of correlation-driven

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.

variance reduction from the cost of heterogeneous judge bias, and a corresponding phase boundary for when aggregation should beat a single judge. This reframes the practical question from "Are diverse panels better?" to "On which evaluation dimensions is the reduction in correlated error large enough to justify the accompanying bias spread?" and yields a falsifiable prediction: reducing inter-judge correlation alone need not monotonically improve panel performance.

Our experiments support this distinction across four independently composed cross-provider panels spanning seven AI providers and four NLG benchmarks (21 dimensions, 2,635 items): every panel reduces inter-judge error correlation relative to a same-provider baseline, yet panel wins do not increase monotonically as correlation falls, and an equal-budget comparison shows a diverse panel is only approximately equivalent to a comparably priced single judge. We detail these findings below.

We make four contributions:

1. **Theory: a diversity-bias decomposition and phase boundary.** We adapt the classical bias-variance-covariance view of ensembles (Krogh & Vedelsby, 1994; Dietterich, 2000; Brown et al., 2005) to first-moment MAE on bounded $[0, 1]$ judge outputs. Proposition 1 separates variance-reduction gain from bias-spread cost and characterizes the regime in which a panel should outperform a single judge.
2. **Evidence that provider diversity reliably reduces correlation, but correlation reduction is not sufficient.** All four independently composed cross-provider panels reduce error correlation on all 21 dimensions relative to the same-provider baseline ($\bar{\rho} = 0.925$), while their dimension-level win rates are not monotonic in $\bar{\rho}$. The decomposition predicts the sign of panel gain on 17/21 dimensions overall and 94% of dimensions satisfying its same-sign-bias assumption (Section 6.1).
3. **Budget- and scale-aware tests of the mechanism.** An equal-budget comparison finds the panel and a comparably priced single judge approximately equivalent (10 versus 11 wins), separating diversity effects from judge quality. At the item level ($N = 4{,}200$), lower error correlation is associated with a small but significant increase in panel wins (Spearman $r = -0.10$, $p < 0.0001$); extending to $M = 5$ judges further increases the variance-reduction coefficient as predicted, subject to the limitation discussed in Section 7.
4. **An actionable deployment rule.** We convert the decomposition into a validation-based, per-dimension rule that chooses between panel and single-judge evaluation (Section 3.4). It improves on the single-judge default on two of four datasets and is intended as a proof of concept for selective panel deployment rather than a tuned production policy.

These results qualify the common intuition that more judge diversity is inherently better. Provider diversity reliably supplies a variance-reduction opportunity; whether that opportunity becomes an accuracy gain depends on the bias it introduces and the strength of the single-judge alternative.

2 Related Work

LLM-as-judge evaluation. Zheng et al. (2023) established LLM-as-judge via MT-Bench and Chatbot Arena (Chiang et al., 2024). Chiang & Lee (2023) and Zeng et al. (2024) studied LLM-human evaluator agreement, and Liu et al. (2023) showed chain-of-thought prompting improves GPT-4 alignment with expert NLG judgments. Kim et al. (2024a) (ICLR 2024), Kim et al. (2024b) (EMNLP 2024), and Zhu et al. (2025) (ICLR 2025) developed open-source rubric-based and fine-tuned judge models. Systematic biases are documented by Shi et al. (2025), Ye et al. (2025), Wang et al. (2024), and Doddapaneni et al. (2024); Lin et al. (2025) and Bai et al. (2024) deploy LLM judges at scale; and Lee et al. (2025) analyze statistical reliability of these pipelines.

Limits and validity of LLM judges. Dorner et al. (2025) (ICLR 2025) establish information-theoretic limits on label efficiency; Calderon et al. (2025) (ACL 2025) give a statistical test for replacing human annotators; Chehbouni et al. (2025) and Maia Polo et al. (2025) (NeurIPS 2025) study reliability, validity, and the human-LLM judgment gap; Jung et al. (2025) (NeurIPS 2025) develop provable human-agreement guarantees; and Adlakha et al. (2024) and Cherian et al. (2024) evaluate correctness and apply conformal prediction to LLM judges.

Multi-judge panels and ensembles. Verga et al. (2024) (PoLL) show cross-provider panels can match a single large judge on pairwise chatbot evaluation, while Kohli (2026) find nine judges may provide only ≈2 independent votes. Turkmen et al. (2026) propose greedy mutual-information selection for budgeted binary-task ensembles; Acharya et al. (2026) (ICML 2026) prove mean aggregation admits unbounded bias under Byzantine corruption; Zhu et al. (2026a) and Zhu et al. (2026b) study calibrator selection and panel expansion; and Madaan et al. (2023) use iterative self-critique as implicit multi-pass aggregation. The classical bias-variance-covariance decomposition underlying our theory is due to Krogh & Vedelsby (1994), surveyed by Brown et al. (2005), Dietterich (2000), and Kuncheva & Whitaker (2003), and unified for general losses by Wood et al. (2023); our contribution is the MAE adaptation with the bias-spread cost correctly signed, applied to LLM judge panels rather than Lakshminarayanan et al. (2017)'s uncertainty-focused deep ensembles.

LLM homogeneity and error correlation. Kim et al. (2025) (ICML 2025) show same-provider models make highly correlated errors. Jiang et al. (2025) (NeurIPS 2025 Best Paper) document a pervasive Artificial Hivemind effect across 70+ frontier LLMs. These results motivate our systematic measurement of provider diversity's effect on inter-judge error correlation.

Multi-attribute NLG evaluation benchmarks. Fabbri et al. (2021) established SummEval (TACL); Mehri & Eskenazi (2020b;a) introduced the USR and FED dialogue datasets, unified by Zhong et al. (2022); Gopalakrishnan et al. (2019); Zhang et al. (2018; 2021) introduced the dialogue datasets we use; Freitag et al. (2021) established multi-attribute MAE via MQM; Fernandes et al. (2023); Adlakha et al. (2024) provide additional multi-attribute evaluation methods.

3 Theoretical Motivation

We adapt the classical ensemble decomposition to motivate the two empirical measurements in Section 5.

3.1 Setup

Item i has true score $y^*_{i,k} \in [0,1]$ on dimension k. Judge j produces $\hat{y}_{i,j,k} \in [0,1]$. Ensemble: $\bar{y}_{i,k} = M^{-1}\sum_j \hat{y}_{i,j,k}$. **Federation gain** $g_k = \min_j \mathrm{MAE}_{j,k} - \mathrm{MAE}(\bar{y}_{\cdot,k}, y^*_{\cdot,k})$ measures how much the panel improves over the best individual judge.

Error decomposition: $e_{i,j,k} = b_{j,k} + \xi_{i,j,k}$, where $b_{j,k}$ is systematic bias and ξ is zero-mean residual noise with variance σ_k^2 and pairwise correlation ρ_k.

Assumption 1 (A1). *$\sigma_{j,k} \approx \sigma_k$ across judges (homogeneous noise variance).*

Assumption 2 (A2). $\Pr(\xi_{i,j,k} > 0) \approx 1/2$ *(approximately symmetric residuals). Holds on 57% of dimensions in our data; see Section 6.1.*

Assumption 3 (A3). *All $b_{j,k}$ share sign on dimension k (same-sign biases; mixed-sign extension in Appendix F).*

3.2 Two-Term Decomposition

Proposition 1 (Federation gain decomposition). *Under Assumptions A1–A3, $M = 3$, sorted bias magnitudes $a \le b \le c$, $s_k = c - a$, $s'_k = b - a$:*

$$g_k = \underbrace{\sigma_{\mathrm{eff},k} \cdot \nu_M(\rho_k)}_{\ge 0,\ \text{variance-reduction gain}} - \underbrace{\frac{s_k + s'_k}{6}}_{\ge 0,\ \text{bias-spread cost}} \tag{1}$$

where $\sigma_{\mathrm{eff},k} = \mathbb{E}|\xi_{i,j,k}|$ and $\nu_M(\rho_k) = 1 - \sqrt{(1 + (M-1)\rho_k)/M}$. Full derivation in Appendix A.

Mechanism. The *variance-reduction gain* $\sigma_{\mathrm{eff},k} \cdot \nu_M(\rho_k)$ is non-negative and strictly decreasing in ρ_k: lower error correlation means more noise cancellation from averaging. The *bias-spread cost* $(s_k + s'_k)/6$ is non-negative and independent of ρ_k: averaging judges with heterogeneous biases worsens the effective ensemble bias relative to the best individual judge. *These two terms pull in opposite directions.*

Remark 1 (The diversity-bias tradeoff)**.** *Provider diversity reduces ρ_k (increasing the gain term) but increases s_k when cross-provider judges have different systematic biases (increasing the cost term). Whether the panel wins depends on which effect dominates, a question that cannot be answered from provider diversity alone.*

Approximation quality. The derivation uses a first-order Taylor approximation of $\mathbb{E}|b + \xi|$ in b, holding when $|b| \ll \sigma_k$; in our data biases and noise are comparable ($|b_j|/\sigma_k \approx 0.5$–$1.0$), so the formula carries magnitude error. A2 holds on 57% of dimensions (12/21; the rest cluster near rubric boundaries, skewing residuals). Despite this, the decomposition predicts the *sign* of federation gain correctly on 17/21 dimensions (81%), validating the two-term structure qualitatively and motivating the two measurements in Section 5: does cross-provider diversity reduce ρ_k, and does lower ρ_k predict more panel wins?

3.3 Phase Boundary

The decomposition implies a phase boundary separating when panels win from when a single judge wins. Let $\delta = \min_j \mathrm{MAE}_{j,k} - \mathrm{MAE}_{\mathrm{large},k}$ be the single large judge's advantage. Panels outperform the large judge when $g_k > \delta$, i.e., when:

$$\sigma_{\mathrm{eff},k} \cdot \nu_M(\rho_k) - \tfrac{s_k + s'_k}{6} > \delta. \tag{2}$$

Under Gaussian noise, this simplifies to $\rho_k < \rho_k^*$ where:

$$\rho_k^* = \tfrac{1}{2}\left[3\left(1 - \tfrac{D_k}{\sigma_{\mathrm{eff},k}}\right)^2 - 1\right], \quad D_k = \delta + \tfrac{s_k + s'_k}{6}. \tag{3}$$

This closed form holds under Gaussian noise assumptions and provides directional guidance: lower correlation, smaller bias spread, and smaller capacity gap all push ρ_k^* higher, making panels more likely to win. We validate the directional prediction empirically in Section 5.6.

3.4 Deployment Rule

From validation estimates, deploy the panel on dimension k iff $\hat{\rho}_k < \hat{\rho}_k^*$, equivalently:

$$\hat{\sigma}_{\mathrm{eff},k} \cdot \hat{\nu}_M(\hat{\rho}_k) > \hat{\delta} + \tfrac{\hat{s}_k + \hat{s}'_k}{6}. \tag{4}$$

Implementation: Appendix G.

4 Experimental Setup

4.1 Datasets

Table 1: Benchmark datasets used in experiments. All human scores normalized to $[0, 1]$.

Dataset	Dims (K)	Items	Task	Ref
SummEval	Coherence, Consistency, Fluency, Relevance (4)	1,600	Summarization	Fabbri et al. (2021)
FED-Turn	Interesting, Engaging, Specific, Relevant, Correct, Sem. Appropriate, Understandable, Fluent, Overall (9)	375	Open dialogue	Mehri & Eskenazi (2020a)
TopicalChat-USR	Naturalness, Coherence, Engagingness, Groundedness (4)	360	Knowledge dial.	Mehri & Eskenazi (2020b)
PersonaChat-USR	Naturalness, Coherence, Engagingness, Groundedness (4)	300	Persona dial.	Mehri & Eskenazi (2020b)
Total	**21 dimensions**	**2,635**	3 task families	

SummEval (Fabbri et al., 2021): 100 CNN/DailyMail articles, 16 summarization systems, expert/crowdworker annotations, 1–5 Likert (`https://huggingface.co/datasets/mteb/summeval`). **FED-Turn** (Mehri & Eskenazi, 2020a): 375 human–chatbot turns, 5 crowdworker

raters/turn (`http://shikib.com/fed_data.json`). **TopicalChat-USR/PersonaChat-USR** (Mehri & Eskenazi, 2020b; Gopalakrishnan et al., 2019; Zhang et al., 2018): expert dialogue-researcher annotations (Zhang et al., 2021), pre-processed by Zhong et al. (2022) (`https://github.com/maszhongming/UniEval`).

4.2 Judge Configurations

Cross-provider panel ($M = 3$). Three judges from three independent AI labs: `claude-haiku-4-5` (Anthropic) (Anthropic, 2025), `Llama-3.3-70B-Instruct` (Meta) (Meta AI, 2024; Grattafiori et al., 2024), and `mistral-small-3.1-24b-instruct` (Mistral AI) (Mistral AI, 2025), both accessed via OpenRouter. These span three independent training pipelines (Anthropic Constitutional AI, Meta open-weight supervised fine-tuning, European instruction tuning), two continents, and open-weight vs. closed paradigms. Following Verga et al. (2024) and Kim et al. (2025), cross-provider diversity is the primary driver of low inter-judge error correlation.

Cost-equivalent single judge. `gpt-4o-mini` (OpenAI) (OpenAI, 2024b), the default choice for practitioners seeking affordable single-judge evaluation; the three-judge panel costs approximately \$0.0011 per item vs. \$0.0002 for a single GPT-4o-mini call (Table 7), 6.8× higher but still 4.4× cheaper than GPT-4o. GPT-4o-mini is deliberately excluded from the panel.

Upper-bound oracle. `gpt-4o` (OpenAI) (OpenAI, 2024a), used to characterize the capacity gap δ and as a frontier reference connecting to the original PoLL comparison (Verga et al., 2024).

Same-provider ablation ($M = 3$). Three instances of `gpt-4o-mini` at temperatures $\{0.0, 0.5, 1.0\}$, run on a 200-item stratified subset per dataset. This isolates the effect of provider diversity from sampling-noise reduction.

Additional panels. Three further panel compositions reuse collected outputs or add two new judges via OpenRouter: (i) *Capability-matched* (all ≈7–12B): Llama-3.1-8B (Meta), Mistral-Nemo (Mistral AI), Qwen-2.5-7B (Alibaba); (ii) *Panel B*: Llama-3.1-8B (Meta), Phi-4 (Microsoft), Nova-Lite (Amazon); (iii) *No-Meta*: Haiku (Anthropic), Mistral-Nemo (Mistral AI), Qwen-2.5-7B (Alibaba), with no Meta model in any slot; (iv) *M=5* on SummEval and FED-Turn: Panel A plus Qwen-2.5-7B and Nova-Micro (Amazon).

All judges use identical prompts requesting per-dimension scores as JSON with values in $[0, 1]$, temperature=0 (except the same-provider ablation). Full prompt templates in Appendix H.

4.3 Cost Breakdown

Full per-judge API pricing and totals are itemized in Appendix B (Table 7). The panel's per-item cost (\$0.00109) exceeds a single GPT-4o-mini call (\$0.00016) by 6.8×, reflecting the sum of three separate API calls, and is 4.4× less than GPT-4o's per-item cost. The correct budget-equivalent framing is therefore: *is a three-judge cross-provider panel worth the additional cost over a single small judge?* We retain GPT-4o-mini as the single-judge baseline because it is the most common practitioner choice for affordable evaluation, and report a truly cost-neutral comparison, which prices a single judge at the panel's per-item budget, separately in Section 5.7.

4.4 Estimation

80/20 random split per dataset: 80% validation for estimating $\hat{\delta}$, $\hat{s}_k$, $\hat{s}'_k$, $\sigma_{\hat{\text{eff}},k}$, $\hat{\rho}_k$; 20% held-out test for deployment-rule evaluation. Error correlation estimated via Pearson correlation of per-judge residuals after bias subtraction. Bootstrap 95% confidence intervals use 1,000 resamples.

Table 2: Mean inter-judge error correlation: cross-provider panel vs. same-provider ensemble (200-item subset). $\Delta\bar{\rho}_k$ is the reduction from diversity. Bootstrap 95% CI half-width ($\pm$, 1,000 resamples) shown per row.

Dataset (K)	**Cross-prov.**	**Same-prov.**	$\Delta\bar{\rho}_k$	**95% CI** ($\pm$)
SummEval (4)	0.717	0.922	-0.205	0.024
FED-Turn (9)	0.669	0.921	-0.252	0.028
TopicalChat-USR (4)	0.813	0.967	-0.154	0.022
PersonaChat-USR (4)	0.764	0.949	-0.185	0.026
Pooled (21)	**0.726**	**0.925**	$\mathbf{-0.199}$	**0.015**

Cross-provider $<$ same-provider on all 21/21 dims (no exceptions).
Distributions do not overlap (cross: 0.553–0.842; same: 0.842–0.974).

5 Results

5.1 Finding 1: Cross-Provider Diversity Consistently Reduces Error Correlation

Cross-provider diversity reduces mean inter-judge error correlation by $\Delta\bar{\rho} = 0.199$ across all 21 evaluation dimensions with no exceptions (Table 2). The two distributions do not overlap: cross-provider rho spans 0.553–0.842; same-provider rho spans 0.842–0.974. This confirms the first part of the tradeoff: provider diversity reliably reduces ρ_k, increasing the variance-reduction gain term in Eq. equation 1.

5.2 Finding 2: Lower Correlation Significantly Predicts More Panel Wins at the Item Level

At $N = 4{,}200$ item-dimension pairs, dimensions with lower inter-judge error correlation show significantly more panel wins:

Table 3: Item-level Spearman correlations ($N = 4{,}200$). Theory (Remark 1) predicts $r < 0$ for both.

Predictor $\rightarrow$ **Outcome**	**Spearman** r	p-value
ρ_k-dim $\rightarrow$ panel beats single (item-level)	-0.102	< 0.0001
ρ_k-dim $\rightarrow$ per-item MAE advantage (cross)	-0.097	< 0.0001
ρ_k-dim $\rightarrow$ cross-over-same advantage	-0.037	0.016

All three associations are in the predicted direction and statistically significant (Table 3). The third row is particularly informative: on lower-rho dimensions, the cross-provider panel has a significantly larger advantage over the same-provider ensemble ($r = -0.037$, $p = 0.016$), isolating the correlation mechanism from all confounds shared between the two panel types. This is the item-level counterpart to Finding 1: not only does provider diversity reduce ρ_k, but lower ρ_k dimensions measurably benefit more from diversity.

5.3 Finding 3: The Tradeoff: Correlation Reduction Does Not Guarantee Panel Wins

Despite the consistent rho reduction in Finding 1, cross-provider panels tie same-provider ensembles at the dimension level:

Table 4: Panel wins vs. GPT-4o-mini (single judge; see Table 7 for cost breakdown). Panel = Haiku + Llama 3.3 + Mistral Small 3.1.

Dataset	K	**Cross-prov. wins**	**Same-prov. wins**	**Mean gain (cross)**
SummEval	4	2/4 (50.0%)	2/4	$+0.0084$
FED-Turn	9	7/9 (77.8%)	6/9	$+0.0354$
TopicalChat-USR	4	3/4 (75.0%)	3/4	$+0.0047$
PersonaChat-USR	4	1/4 (25.0%)	2/4	-0.0052
Pooled	**21**	**13/21 (61.9%)**	**13/21 (61.9%)**	$+0.0167$

Cross-provider panels win on 13/21 dimensions against GPT-4o-mini; same-provider ensembles also win on 13/21. Both configurations tie despite the large rho difference (0.726 vs. 0.925). This pattern

is consistent with the diversity-bias tradeoff (Remark 1): cross-provider panels gain from lower ρ_k (variance-reduction gain increases) but pay a higher bias-spread cost, with mean cross-provider bias spread $\bar{s}_k = 0.097$ versus $\bar{s}_k \approx 0$ for same-provider, where biases are nearly identical. We read this as evidence consistent with the two effects partially offsetting. The item-level analysis (Section 6.1) is statistically better powered and points the same direction ($r = -0.10$, $p < 0.0001$).

5.4 Finding 4: Rho Reduction Generalizes Across Panel Compositions

To test whether the rho reduction is a robust property of cross-provider diversity rather than an artifact of Panel A's specific composition, we run two additional panels:

Capability-matched panel. Three judges at similar capability scale from different labs: `Llama-3.1-8B` (Meta, 8B), `Mistral-Nemo` (Mistral, 12B), and `Qwen-2.5-7B` (Alibaba, 7B). This panel controls for model size, isolating provider identity as the source of correlation reduction. Mean rho: $\bar{\rho} = 0.633$, below same-provider (0.925) on all 21/21 dimensions. The reduction ($\Delta\bar{\rho} = 0.292$) is larger than Panel A's ($\Delta\bar{\rho} = 0.199$), confirming that provider diversity, not capability spread, drives the rho reduction.

Panel B: independently-composed panel. Three judges from three different providers not in Panel A: `Llama-3.1-8B` (Meta), `Phi-4` (Microsoft), `Nova-Lite` (Amazon). Mean rho: $\bar{\rho} = 0.493$, below same-provider on 21/21 dimensions (range: 0.158–0.670). Panel B wins on 15/21 dimensions (71.4%) vs. GPT-4o-mini, compared to 13/21 (61.9%) for Panel A.

Table 5: Rho reduction and panel wins across all four cross-provider panel compositions vs. same-provider baseline ($\bar{\rho} = 0.925$). No-Meta panel uses no shared provider slot with Panel B.

Panel	Providers	$\bar{\rho}$	$\Delta\bar{\rho}$	Wins/21
Same-provider (baseline)	GPT-4o-mini ×3	0.925	—	13/21
Panel A (original)	Anthropic + Meta + Mistral	0.726	−0.199	13/21
Capability-matched	Meta + Mistral + Alibaba	0.633	−0.292	12/21
Panel B	Meta + Microsoft + Amazon	0.493	−0.432	15/21
No-Meta panel	Anthropic + Mistral + Alibaba	0.606	−0.319	12/21

All four panels reduce rho on 21/21 dims;
no shared provider slot between No-Meta and Panel B.

To close Limitation 8 (shared Meta slot), we run a fourth panel with no Meta model: Haiku (Anthropic) + Mistral-Nemo (Mistral AI) + Qwen-2.5-7B (Alibaba), reusing already-collected judge outputs. This no-Meta panel achieves $\bar{\rho} = 0.606$ on 21/21 dimensions below 0.925, with 12/21 wins vs. GPT-4o-mini. Across all four cross-provider panels, spanning seven distinct AI providers, both open-weight and closed models, and both capability- matched and frontier-scale judges, rho reduction below the same-provider baseline holds on every dimension tested. Win rates are not monotonic in $\bar{\rho}$ (Table 5); this non-monotonicity is addressed in Section 6.4.

5.5 Finding 5: Task Structure Modulates the Tradeoff

Panel gains are strongly task-dependent: FED-Turn (open dialogue, 9 dims) shows the largest panel advantage (7/9 wins, mean gain $+0.035$), while SummEval (summarization, 4 dims) shows the smallest (2/4 wins, mean gain $+0.008$). This is consistent with the phase boundary interpretation (Eq. equation 3): on summarization tasks, GPT-4o-mini has a larger capacity advantage over small models (δ large), raising the threshold D_k and making ρ_k^* smaller, so that observed ρ_k exceeds ρ_k^* and the single judge wins. On dialogue tasks, the capacity gap shrinks, bringing ρ_k^* within reach of observed ρ_k.

5.6 Phase Boundary Directional Validation

Binning dimensions by ρ_k quartile, the highest-rho quartile shows the lowest win rate (40%) and the lowest-rho quartile shows the highest (67%), directionally consistent with the phase boundary

prediction, though with only $n = 5$–6 dimensions per quartile the binning is not monotonic (full table and discussion in Appendix D). The item-level result in Table 3 remains the primary statistical evidence for this relationship.

5.7 Cost-Neutral Comparison

The panel ($0.00109/item) costs 6.8× more than GPT-4o-mini ($0.00016/item). A cost-neutral single-judge baseline spends the panel's total budget on one judge; at ≈ $0.001/item, Claude Haiku 4.5 ($0.00092/item) is the closest available model (92% of panel budget), requiring no additional API calls since Haiku outputs are already collected. At equal budget, Haiku alone outperforms the panel on 11/21 dimensions while the panel beats Haiku on 10/21 (Table 6).

Table 6: Cost-neutral comparison: panel vs. single Haiku judge at approximately equal per-item cost.

Configuration	Cost/item	Wins/21	vs GPT-4o-mini
Panel A (3 judges)	$0.00109	10/21	13/21
Single Haiku (Anthropic)	$0.00092	11/21	11/21
GPT-4o-mini (baseline)	$0.00016	—	—

At equal budget, single Haiku and the panel are approximately equivalent (11 vs. 10 wins).

This reframes the panel's advantage: the 13/21 win rate over GPT-4o-mini (Section 5.3) reflects partly Haiku's individual quality and partly diversity, inseparable from that comparison alone; at equal budget, diversity adds little further value, so cost-neutral practitioners should not expect a panel to reliably beat their best available single judge. A panel is still worth its complexity when (i) the strongest individual judge is not known in advance, so diversity hedges against picking the wrong one, or (ii) the deployment rule (Section 5.8) routes each dimension to whichever configuration validation data favors, rather than committing to one configuration paper-wide. This is a more defensible claim than "panels beat single judges," and the one this paper's contributions actually support.

5.8 Deployment Rule

The validation-driven deployment rule (Eq. equation 4) improves over both naive defaults (Always-Panel, Always-Single) on FED-Turn and TopicalChat-USR, the two datasets with the largest panel gains, and moves in the correct direction on SummEval; on PersonaChat-USR it matches Always-Panel but is outperformed by Always-Single, reflecting noisier validation estimates at $n_{\text{val}} = 240$. Full per-dataset MAE by policy is in Appendix E (Table 9); absolute margins are small on all four datasets and the rule is a proof of concept, not a settled tool.

6 Analysis

6.1 Assumption Diagnostics

Per-dimension pass rates: (A1) variance homogeneity: 20/21 (95%); (A2) symmetric noise: 12/21 (57%, failing on boundary-heavy dimensions where judges cluster near rubric ceilings (Fabbri et al., 2021; Mehri & Eskenazi, 2020a)); (A3) same-sign biases: 17/21 (81%). Panel win rates do not differ significantly between passing and failing dimensions (Lee et al., 2025). Sign accuracy separately: 16/17 (94%) on A3-compliant dimensions, 1/4 (25%) on A3-violating dimensions, confirming A3 compliance is the primary driver of the 81% headline figure.

6.2 Bias Spread as the Limiting Factor

Comparing pairwise rho against pairwise win rate for all three judge pairs (Haiku–Llama, Haiku–Mistral, Llama–Mistral), the haiku–mistral pair wins most (16/21 = 76%) despite not having the lowest rho (0.758 vs. llama–mistral's 0.693), because it also has lower bias spread; both terms in Eq. equation 1 jointly determine outcomes (Kim et al., 2025; Jiang et al., 2025). The capability-matched panel (Section 5.4) confirms bias spread persists even at $\bar{\rho} = 0.633$, despite similar model sizes.

6.3 M=5 PANEL AND THE ν_M FORMULA

We run $M = 5$ panels on both SummEval and FED-Turn, adding Qwen-2.5-7B and Nova-Micro to the original three Panel A judges. ν_M rises with M on both datasets (SummEval: 0.119→0.167; FED-Turn: 0.078→0.128), consistent with the formula. The fixed-correlation decomposition (Appendix C) attributes 52% of the SummEval increase to M alone and 32% on FED-Turn, a less clean separation since the added judges also lower FED-Turn's correlation substantially. Wins move with the prediction where bias spread does not dominate (SummEval: 2/4→3/4; FED-Turn unchanged at 7/9, already near ceiling at $M = 3$).

6.4 SAME-PROVIDER ABLATION

The same-provider ablation (3× GPT-4o-mini, temperatures $\{0.0, 0.5, 1.0\}$) achieves mean rho 0.925 vs. 0.726 for the cross-provider panel, yet ties at 13/21 dimension wins; bias spread s_k rises from ≈0 (same model) to 0.142 (heterogeneous providers), consistent with the tradeoff offsetting the correlation gain at the win-count level. The capability-matched panel (Section 5.4), controlled at 7–12B parameters, achieves an even lower $\bar{\rho} = 0.633$ than Panel A, confirming provider identity, not capability spread, drives correlation reduction, yet wins only 12/21 dimensions, one fewer than both Panel A and the same-provider baseline, a gap a bootstrap diagnostic ($B = 2{,}000$ resamples) finds within the noise floor ($p = 1.00$, 95% CI $[1.0, 1.0]$) under both residual-shuffle and bias-shuffle nulls.

7 DISCUSSION

Why prior work found mixed results, and the Hivemind ceiling. Verga et al. (2024) found cross-provider panels match frontier judges; Kohli (2026) found nine judges deliver only ≈2 effective votes; both are consistent with our tradeoff, since correlation reduction dominates when bias spread is low, offset when it is not. Separately, Jiang et al. (2025)'s finding that 70+ frontier LLMs converge through shared RLHF norms implies $\rho_k \to 1$ and $\nu_M(\rho_k) \to 0$ as models homogenize; cross-provider diversity partially counters this (0.726 vs. 0.925), but caps how much diversity is achievable.

When to use cross-provider panels in practice. Panels suit subjective dialogue tasks with competitive, low-cost small models and similarly-calibrated providers (low s_k); single large judges suit tasks with large capacity gaps (e.g. summarization) or high provider bias heterogeneity.

Limitations. (1) The Taylor approximation is qualitative, with error growing when $|b_j|/\sigma_k \approx 0.5$–$1.0$, and (2) the Gaussian closed form for ρ_k^* (Eq. equation 3) only approximates the boundary under non-Gaussian distributions. (3) Dimension-level analysis ($N = 21$) is underpowered relative to the item-level result ($N = 4{,}200$). (4) The $M = 5$ decomposition confounds M with a concurrent rho shift; a fully controlled fixed-ρ_k test remains untested. (5) MAE-to-reference differs from human preference quality (Fabbri et al., 2021; Freitag et al., 2021). (6) The capability-matched panel wins one fewer dimension than the same-provider baseline despite lower $\bar{\rho}$, within the bootstrap noise floor. (7) The panel costs 6.8× more per item than GPT-4o-mini; at equal budget, Haiku alone and the panel are roughly equivalent (Section 5.7).

8 CONCLUSION

Diversity alone does not decide whether an LLM evaluation pipeline should use a panel or a single strong judge. Cross-provider panels consistently reduce inter-judge error correlation, but this variance benefit competes with heterogeneous systematic bias, a tradeoff we formalized with a two-term decomposition and phase boundary whose sign prediction is correct on 94% of dimensions satisfying its same-sign-bias assumption (81% overall). Across four cross-provider panels spanning seven providers, lower correlation was a robust consequence of diversity, though it did not guarantee better evaluation.

This matters in practice: at equal budget, the panel and a comparably priced single judge are approximately equivalent (10 versus 11 wins), so gains over a weaker baseline can partly reflect judge quality rather than diversity itself. The lesson is to measure the tradeoff and route evaluation accordingly, which our deployment rule does per dimension.

Reproducibility Statement

All four datasets are publicly available with permanent URLs provided in Section 4. Judge model versions, API access dates, and prompt templates are documented in Section 4.2 and Appendix H. Analysis code, preprocessing scripts, and the deployment rule implementation are submitted as anonymized supplementary material. All quantitative claims are derived from the experimental data described above; no numbers are taken from prior work without citation.

AI Use Statement

Claude Sonnet 4.6 (Anthropic) was used to aid LaTeX formatting and polish writing, and to assist with literature search organization. All theoretical derivations, experimental designs, data collection, analysis, interpretations, and scientific claims were developed and verified by the authors. We take full responsibility for the content of this paper.

Ethics Statement

This work uses only publicly available NLG evaluation benchmarks with no personally identifiable information. API calls used public commercial endpoints under standard terms of service. MAE-to-reference should not substitute for human evaluation in high-stakes settings. Judge models may exhibit biases affecting evaluation quality; our framework measures this empirically via the bias-spread term.

References

Anish Acharya, Kris W Pan, and Brian Verkhovsky. RopoLL: Robust panel of LLM judges. In *ICML 2026 Statistical Frameworks for Uncertainty in Agentic Systems*, 2026. URL https://openreview.net/forum?id=Rvj0jSb86u.

Vaibhav Adlakha, Parishad BehnamGhader, Xing Han Lu, Nicholas Meade, and Siva Reddy. Evaluating correctness and faithfulness of instruction-following models for question answering. *Transactions of the Association for Computational Linguistics*, 12:681–699, 2024. doi: 10.1162/tacl_a_00667. URL https://aclanthology.org/2024.tacl-1.38/.

Anthropic. Claude haiku 4.5: Fast and efficient small model, 2025. URL https://www.anthropic.com/claude/haiku. Model API: claude-haiku-4-5. Released October 2025.

Ge Bai, Jie Liu, Xingyuan Bu, Yancheng He, Jiaheng Liu, Zhanhui Zhou, Zhuoran Lin, Wenbo Su, Tiezheng Ge, Bo Zheng, and Wanli Ouyang. MT-bench-101: A fine-grained benchmark for evaluating large language models in multi-turn dialogues. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7421–7454. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.401. URL https://aclanthology.org/2024.acl-long.401/.

Gavin Brown, Jeremy Wyatt, Rachel Harris, and Xin Yao. Diversity creation methods: a survey and categorisation. *Information fusion*, 6(1):5–20, 2005. URL https://doi.org/10.1016/j.inffus.2004.04.004.

Nitay Calderon, Roi Reichart, and Rotem Dror. The alternative annotator test for LLM-as-a-judge: How to statistically justify replacing human annotators with LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16051–16081. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.782. URL https://aclanthology.org/2025.acl-long.782/.

Khaoula Chehbouni, Mohammed Haddou, Jackie CK Cheung, and Golnoosh Farnadi. Neither valid nor reliable? investigating the use of llms as judges. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference. Curran Associates, Inc., 2025. doi: 10.52202/085713-3029. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/829e8f32d76a6248815bf5b01633811d-Paper-Position_Paper_Track.pdf.

John J. Cherian, Isaac Gibbs, and Emmanuel J. Candès. Large language model validity via enhanced conformal prediction methods. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 114812–114842. Curran Associates, Inc., 2024. doi: 10.52202/079017-3645. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/d02ff1aeaa5c268dc34790dd1ad21526-Paper-Conference.pdf.

Cheng-Han Chiang and Hung-yi Lee. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15607–15631. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.870. URL https://aclanthology.org/2023.acl-long.870/.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios N. Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: An open platform for evaluating LLMs by human preference. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024. doi: 10.48550/arXiv.2403.04132.

Thomas G. Dietterich. Ensemble methods in machine learning. In *Multiple Classifier Systems*, pp. 1–15, 2000.

Sumanth Doddapaneni, Mohammed Safi Ur Rahman Khan, Sshubam Verma, and Mitesh M Khapra. Finding blind spots in evaluator LLMs with interpretable checklists. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 16279–16309. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.911. URL https://aclanthology.org/2024.emnlp-main.911/.

Florian Eddie Dorner, Vivian Nastl, and Moritz Hardt. Limits to scalable evaluation at the frontier: Llm as judge won,Äôt beat twice the data. In *International Conference on Learning Representations*, volume 2025, pp. 26467–26491, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/4264ee4376776907c0b87ed70b959585-Paper-Conference.pdf.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409, 2021. doi: 10.1162/tacl_a_00373.

Patrick Fernandes, Daniel Deutsch, Mara Finkelstein, Parker Riley, André Martins, Graham Neubig, Ankush Garg, Jonathan Clark, Markus Freitag, and Orhan Firat. The devil is in the errors: Leveraging large language models for fine-grained machine translation evaluation. In *Proceedings of the Eighth Conference on Machine Translation*, pp. 1066–1083. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.wmt-1.100. URL https://aclanthology.org/2023.wmt-1.100/.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474, 2021. doi: 10.1162/tacl_a_00437. URL https://aclanthology.org/2021.tacl-1.87/.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anushree Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations. In *Proceedings of Interspeech*, pp. 1891–1895, 2019. doi: 10.21437/Interspeech.2019-3079.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. doi: 10.48550/arXiv.2407.21783.

Liwei Jiang, Yuanjun Chai, Margaret Li, Mickel Liu, Raymond Fok, Nouha Dziri, Yulia Tsvetkov, Maarten Sap, and Yejin Choi. Artificial hivemind: The open-ended homogeneity of language models (and beyond). In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, 2025. doi: 10.52202/085713-2732.

URL https://proceedings.neurips.cc/paper_files/paper/2025/file/754d5a526a5ee5a47220664a0eb92751-Paper-Datasets_and_Benchmarks_Track.pdf.

Jaehun Jung, Faeze Brahman, and Yejin Choi. Trust or escalate: Llm judges with provable guarantees for human agreement. In *International Conference on Learning Representations*, volume 2025, pp. 3101–3125, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/08dabd5345b37fffcbe335bd578b15a0-Paper-Conference.pdf.

Elliot Kim, Avi Garg, Kenny Peng, and Nikhil Garg. Correlated errors in large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 30038–30066. PMLR, 2025. doi: 10.21427/kim25e.

Seungone Kim, Jay Shin, yejin cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, S Shin, Ryan, Sungdong Kim, James Thorne, and Minjoon Seo. Prometheus: Inducing fine-grained evaluation capability in language models. In *International Conference on Learning Representations*, volume 2024, pp. 29927–29962, 2024a. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/803485352e61e3ebf41221e4776c9fd4-Paper-Conference.pdf.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. Prometheus 2: An open source language model specialized in evaluating other language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 4334–4353. Association for Computational Linguistics, 2024b. doi: 10.18653/v1/2024.emnlp-main.248. URL https://aclanthology.org/2024.emnlp-main.248/.

Guneet Kohli. Nine judges, two effective votes: Correlated errors undermine LLM evaluation panels. *arXiv.org*, 2026. doi: 10.48550/arXiv.2605.29800.

Anders Krogh and Jesper Vedelsby. Neural network ensembles, cross validation, and active learning. In *Advances in Neural Information Processing Systems*, volume 7, 1994. URL https://proceedings.neurips.cc/paper_files/paper/1994/file/b8c37e33defde51cf91e1e03e51657da-Paper.pdf.

Ludmila I Kuncheva and Christopher J Whitaker. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine learning*, 51(2):181–207, 2003. doi: 10.1023/A:1022859003006.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/9ef2ed4b7fd2c810847ffa5fa85bce38-Paper.pdf.

Chungpa Lee, Thomas Zeng, Jongwon Jeong, Jy-yong Sohn, and Kangwook Lee. How to correctly report llm-as-a-judge evaluations. *arXiv preprint arXiv:2511.21140*, 2025. doi: 10.48550/arXiv.2511.21140.

Bill Yuchen Lin, Yuntian Deng, Khyathi Chandu, Abhilasha Ravichander, Valentina Pyatkin, Nouha Dziri, Ronan Le Bras, and Yejin Choi. Wildbench: Benchmarking llms with challenging tasks from real users in the wild. In *International Conference on Learning Representations*, volume 2025, pp. 47852–47870, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/771155abaae744e08576f1f3b4b7ac0d-Paper-Conference.pdf.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511–2522. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.153. URL https://aclanthology.org/2023.emnlp-main.153/.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46534–46594. Curran Associates, Inc., 2023. doi: 10.52202/075280-2019. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/91edff07232fb1b55a505a9e9f6c0ff3-Paper-Conference.pdf.

Felipe Maia Polo, Xinhe Wang, Mikhail Yurochkin, Gongjun Xu, Moulinath Banerjee, and Yuekai Sun. Bridging human and llm judgments: Understanding and narrowing the gap. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 15049–15100. Curran Associates, Inc., 2025. doi: 10.52202/085713-0507. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/160f484617c9e3ce84332d8cbc8ba62b-Paper-Conference.pdf.

Shikib Mehri and Maxine Eskenazi. Unsupervised evaluation of interactive dialog with DialoGPT. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 225–235. Association for Computational Linguistics, 2020a. doi: 10.18653/v1/2020.sigdial-1.28. URL https://aclanthology.org/2020.sigdial-1.28/.

Shikib Mehri and Maxine Eskenazi. USR: An unsupervised and reference free evaluation metric for dialog generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 681–707, Online, 2020b. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.64. URL https://aclanthology.org/2020.acl-main.64/.

Meta AI. Llama 3.3: Meta's 70b instruction-tuned model, 2024. URL https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct. Model accessed via OpenRouter. Released December 2024.

Mistral AI. Mistral Small 3.1 24B: Efficient multilingual model, 2025. URL https://mistral.ai/news/mistral-small-3-1/. Model API: mistral-small-3.1-24b-instruct. Released March 2025. Accessed via OpenRouter.

OpenAI. GPT-4o: OpenAI's flagship multimodal model, 2024a. URL https://openai.com/index/hello-gpt-4o/. Model API: gpt-4o. Released May 2024.

OpenAI. GPT-4o mini: Advancing cost-efficient intelligence, 2024b. URL https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/. Model API: gpt-4o-mini. Released July 2024.

Lin Shi, Chiyu Ma, Wenhua Liang, Xingjian Diao, Weicheng Ma, and Soroush Vosoughi. Judging the judges: A systematic study of position bias in LLM-as-a-judge. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 292–314. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.ijcnlp-long.18. URL https://aclanthology.org/2025.ijcnlp-long.18/.

Yigit Turkmen, Baturalp Buyukates, and Melih Bastopcu. Don't always pick the highest-performing model: An information theoretic view of LLM ensemble selection. *arXiv.org*, 2026. doi: 0.48550/arXiv.2602.08003.

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. doi: 10.48550/arXiv.2404.18796.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018. ISBN 9781108415194. doi: 10.1017/9781108231596.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 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. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.511. URL https://aclanthology.org/2024.acl-long.511/.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837, 2022. doi: 10.52202/068431-1800. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf.

Danny Wood, Tingting Mu, Andrew M Webb, Henry WJ Reeve, Mikel Lujan, and Gavin Brown. A unified theory of diversity in ensemble learning. *Journal of machine learning research*, 24(359): 1–49, 2023. URL http://jmlr.org/papers/v24/23-0041.html.

Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, Nitesh Chawla, and Xiangliang Zhang. Justice or prejudice? quantifying biases in llm-as-a-judge. In *International Conference on Learning Representations*, volume 2025, pp. 102351–102390, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/fdca08d371e4b6c031397909e20043bd-Paper-Conference.pdf.

Zhiyuan Zeng, Jiatong Yu, Tianyu Gao, Yu Meng, Tanya Goyal, and Danqi Chen. Evaluating large language models at evaluating instruction following. In *International Conference on Learning Representations*, volume 2024, pp. 40193–40219, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/afc8b034823271816d14f7c1aefe1dff-Paper-Conference.pdf.

Chen Zhang, João Sedoc, Luis Fernando D'Haro, Rafael Banchs, and Alexander Rudnicky. Automatic evaluation and moderation of open-domain dialogue systems. *arXiv preprint arXiv:2111.02110*, 2021. doi: 10.48550/arXiv.2111.02110.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2204–2213. Association for Computational Linguistics, 2018. doi: 10.18653/v1/P18-1205. URL https://aclanthology.org/P18-1205/.

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. doi: 10.52202/075280-2020.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. Towards a unified multi-dimensional evaluator for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 2023–2038. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.emnlp-main.131. URL https://aclanthology.org/2022.emnlp-main.131/.

Bin Zhu, Yi Xie, and Yanghui Rao. A finite-calibration regime map for llm judge panels. *arXiv preprint arXiv:2606.01034*, 2026a. doi: 10.48550/arXiv.2606.01034.

Bin Zhu, Yi Xie, and Yanghui Rao. Stopping and routing llm judge panels. *arXiv preprint arXiv:2608.19802*, 2026b. doi: 10.48550/arXiv.2608.19802.

Lianghui Zhu, Xinggang Wang, and Xinlong Wang. Judgelm: Fine-tuned large language models are scalable judges. In *International Conference on Learning Representations*, volume 2025, pp. 51257–51296, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/7f8f73134e253845a8f82983219a8452-Paper-Conference.pdf.

A Full Derivation of Proposition 1

Step 1: Per-judge MAE. Under A2, $\Pr(\xi > 0) \approx 1/2$, so $f(b) = \mathbb{E}|b + \xi|$ satisfies $f'(0) \approx 0$ and $f(0) = \sigma_{\text{eff},k}$. First-order Taylor: $\text{MAE}_{j,k} \approx \sigma_{\text{eff},k} + |b_{j,k}|/2$. Best judge: $\min_j \text{MAE}_{j,k} \approx \sigma_{\text{eff},k} + a/2$.

Step 2: Ensemble MAE. Ensemble residual variance: $\text{Var}(\bar{\xi}_{i,k}) = \sigma_k^2(1+(M-1)\rho_k)/M = \sigma_k^2 Q_M$. Under A2: $\mathbb{E}|\bar{\xi}_{i,k}| \approx \sigma_{\text{eff},k}\sqrt{Q_M(\rho_k)}$. Same-sign biases (A3): $|\bar{b}_k| = (a+b+c)/3$. Ensemble MAE: $\sigma_{\text{eff},k}\sqrt{Q_M} + (a+b+c)/6$.

Step 3: Gain. $g_k = [\sigma_{\text{eff},k} + a/2] - [\sigma_{\text{eff},k}\sqrt{Q_M} + (a+b+c)/6] = \sigma_{\text{eff},k}(1-\sqrt{Q_M}) - (s'_k + s_k)/6$. □

Exactness note. Step 1 uses Taylor approximation; it is exact when ξ is exactly symmetric (A2 holds exactly) and becomes approximate when A2 holds partially. The bias-spread cost $(s_k + s'_k)/6$ is exact for $M = 3$ same-sign biases regardless of noise distribution. The sub-Gaussian concentration bound for the regression estimator follows Vershynin (2018), Theorem 2.6.3.

B Cost Breakdown

Table 7: API pricing and total experiment cost per judge (2,635 items, ≈650 input tokens and 100 output tokens per call). Panel total is the sum of the three cross-provider judges.

Judge	$/1M in	$/1M out	$/item	Total (2,635)
Claude Haiku 4.5 (Anthropic)	$0.80	$4.00	$0.00092	$2.42
Llama 3.3-70B (OpenRouter)	$0.10	$0.10	$0.00007	$0.20
Mistral Small 3.1 (OpenRouter)	$0.10	$0.30	$0.00010	$0.25
Panel (3 judges)			**$0.00109**	**$2.87**
GPT-4o-mini / single judge (OpenAI)	$0.150	$0.600	$0.00016	$0.42
GPT-4o / oracle (OpenAI)	$5.00	$15.00	$0.00475	$12.52

Panel costs 6.8× more per item than a single GPT-4o-mini call, but 4.4× less than GPT-4o. Pricing as of September 2026 from provider public pricing pages.

C M=5 Fixed-Correlation Decomposition

To isolate the M-term of $\nu_M(\rho_k) = 1 - \sqrt{(1+(M-1)\rho_k)/M}$ from the correlation shift observed when moving from the $M = 3$ Panel A judges to the $M = 5$ panel on SummEval, we evaluate the formula at the $M = 3$ correlation ($\bar{\rho} = 0.665$) with M set to 5: $\nu_5(0.665) = 1 - \sqrt{(1 + 4 \times 0.665)/5} = 1 - \sqrt{3.66/5} = 1 - \sqrt{0.732} \approx 1 - 0.8556 \approx 0.144$. This isolates the pure M-effect predicted by the formula: moving from $\nu_3(0.665) = 0.119$ to $\nu_5(0.665) \approx 0.144$ is attributable to M alone, holding correlation fixed. The remaining move from 0.144 to the observed $\nu_5(0.620) \approx 0.167$ is attributable to the correlation shift from 0.665 to 0.620 that resulted from adding two judges with their own correlation profile, not to the M term. Of the total observed increase on SummEval (0.119 to 0.167, i.e. 0.048), roughly 0.025 (52%) is attributable to M and roughly 0.023 (48%) to the correlation shift.

On FED-Turn ($M = 3$: mean $\rho = 0.776$, $\nu_3 = 0.078$; $M = 5$: mean $\rho = 0.701$, $\nu_5 = 0.128$): evaluating the formula at $M = 5$ holding ρ fixed at 0.776 gives $\nu_5(0.776) = 1 - \sqrt{(1 + 4 \times 0.776)/5} \approx 0.094$. Of the total increase 0.050, roughly 0.016 (32%) is attributable to M alone and 0.034 (68%) to the correlation shift. Win rates are unchanged (7/9 at both $M = 3$ and $M = 5$), consistent with FED-Turn already showing strong panel gains at $M = 3$ where the single-judge margin is large. A controlled test isolating the M term alone would need to hold ρ_k fixed while varying M, which we leave to future work.

D Phase Boundary Quartile Table

The Q2>Q1 non-monotonicity is attributable to small per-quartile sample size ($n = 5$–6) and the confound that low-rho dimensions have higher bias spread (which partially offsets the gain).

Table 8: Panel win rate by error correlation quartile. Theory predicts monotonically decreasing win rate.

Quartile	Mean ρ_k	Win rate	Mean gain	n
Q1 (low, 0.55–0.68)	0.620	66.7%	$+0.024$	6
Q2 (0.68–0.76)	0.708	80.0%	$+0.046$	5
Q3 (0.76–0.80)	0.779	60.0%	$+0.010$	5
Q4 (high, 0.80–0.84)	0.823	40.0%	-0.004	5

Q4 (40%) $<$ Q1 (67%): highest-rho quartile shows fewest wins, consistent with theory.

The overall direction, high-rho dimensions lose more often, is consistent with the phase boundary prediction, but this binning is underpowered on its own; see Table 3 in the main text for the item-level result, which is the stronger evidence.

E Deployment Rule: Per-Dataset MAE

Table 9: Mean test-set MAE by deployment policy. Lower is better. Rule uses Eq. equation 4 on the validation split.

Dataset	A-Panel	A-Single	Same-Prov	Rule	Oracle
SummEval	0.159	0.151	0.152	0.156	0.150
FED-Turn	0.389	0.397	0.396	0.385	0.380
TopicalChat-USR	0.276	0.276	0.275	0.269	0.267
PersonaChat-USR	0.265	0.261	0.261	0.265	0.247

The rule improves over both naive defaults on FED-Turn and TopicalChat-USR, the two datasets with the largest panel gains, and moves in the correct direction on SummEval (0.156, between Always-Single's 0.151 and Always-Panel's 0.159, closer to the single-judge optimum). On PersonaChat-USR the rule matches Always-Panel (0.265), outperformed by Always-Single (0.261), reflecting noisier validation estimates at $n_{\text{val}} = 240$ items.

F Mixed-Sign Bias Extension

When biases have mixed signs: let $B_+ = \sum_{j:b_{j,k}>0} b_{j,k}$, $B_- = \sum_{j:b_{j,k}<0} |b_{j,k}|$. General form: $g_k = \sigma_{\text{eff},k} \cdot \nu_M(\rho_k) + a/2 - |B_+ - B_-|/(2M)$. When $B_+ \approx B_-$, bias cancellation adds to the gain beyond variance reduction. Eq. equation 1 is the same-sign special case ($B_- = 0$).

G Deployment Rule Code

```
import numpy as np

def phase_boundary(mae_panel_judges,
                   mae_single, sigma_eff, M=3):
    mae_j = np.sort(mae_panel_judges)
    a, b, c = mae_j
    delta = mae_j.min() - mae_single
    D = delta + (c - a + b - a) / 6
    if D >= sigma_eff:
        return -1.0   # panel never wins
    return 0.5 * (3 * (1 - D/sigma_eff)**2 - 1)

def deploy_panel(mae_panel_judges, mae_single,
                 rho_k, sigma_eff):
    rho_star = phase_boundary(
        mae_panel_judges, mae_single, sigma_eff)
    return rho_k < rho_star
```

H JUDGE PROMPT TEMPLATES

All judges receive the same prompt per task type. For summarization evaluation:

```
You are an expert NLG evaluator.  Score the following
summary on each dimension.  Each score must be a float
between 0.0 (worst) and 1.0 (best).  Return ONLY a
JSON object, no explanation.
ARTICLE: {article[:1500]}
SUMMARY: {summary[:500]}
{"coherence":  <float>, "consistency":  <float>, ...}
```

For dialogue evaluation, "ARTICLE" and "SUMMARY" are replaced by "CONTEXT" and "RESPONSE". All judges use temperature=0 except the same-provider ablation, following standard practice (Liu et al., 2023; Wei et al., 2022; Bai et al., 2024; Kim et al., 2024b).