

Beyond the Illusion of Consensus: From Surface Heuristics to Knowledge-Grounded Evaluation in LLM-as-a-Judge

Mingyang Song, Mao Zheng

Foundation Model Department, Tencent, China

nickmysong@tencent.com

Abstract

Agreement among LLM judges is widely used as evidence that automated evaluation is reliable. Yet consensus may be stable without reflecting a shared understanding of quality. We call this **Evaluation Illusion**: judges can produce detailed critiques while relying on overlapping surface cues or a common evaluation instrument. In a fully crossed baseline study, three frontier judges conduct 105,600 evaluations of 32 models. First, for the two highest-scoring models, moving from a shared five-dimension rubric package to independently generated rubrics reduces agreement by 62%, a joint contrast we term the **Rubric Commensurability Problem**. Second, nearly identical model-level rankings coexist with substantially weaker criterion-level agreement under non-equivalent metrics and sample sets, yielding the **Resolution Paradox**. We further introduce MERG, a four-stage protocol that elicits task knowledge and potential biases before scoring. In a controlled comparison on the same models at $t=0.0$, agreement decreases under independently generated MERG rubrics. These results do not establish that MERG scores are more accurate; they show that consensus depends on the evaluation instrument and analytical resolution at which it is measured.

1 Introduction

Large language models (LLMs) now rank models, label preferences, and supply rewards in alignment pipelines (Zheng et al., 2023; Chiang et al., 2024; Lee et al., 2023). Agreement with humans is often used as a proxy for validity, while agreement among judges is treated as evidence of consistency. Results from MT-Bench, G-Eval, and AlpacaEval helped establish both practices (Zheng et al., 2023; Liu et al., 2023; Li et al., 2023; Chiang et al., 2024), and rubric-based protocols made automated evaluation appear still more systematic (Ye et al., 2024; Kim et al., 2024a,b; Rao and Callison-Burch, 2026;

Wu et al., 2025). But agreement alone cannot establish that judges rely on substantive evidence rather than merely similar signals.

This gap motivates **Evaluation Illusion**: apparently expert judgments can converge even when their scores remain sensitive to a shared instrument or overlapping surface cues. We call the resulting consensus a **Shared Illusion** when it is statistically stable yet epistemically shallow. Our quantitative analyses measure sensitivity to rubric design and aggregation; our case studies show how fluent critiques can emphasize formatting, confident tone, and structural polish. Figure 1 illustrates the distinction. Two displayed baseline judges assign very high scores to a polished pitch for a Chinese online-education platform while overlooking a placeholder license, an unsupported subsidy claim, and the absence of a clear compliance pathway. The agreement is real. Its meaning is not.

We introduce **MERG** (Metacognitive Enhanced Rubric Generation) to probe whether such consensus survives explicit deliberation. MERG adds three pre-scoring stages in which each judge articulates task-relevant knowledge, anticipates surface biases, and generates a rubric before conducting a final evidence-based evaluation. The design is inspired by the distinction between fast heuristic and slower deliberative processing (Kahneman, 2011). Here, *knowledge grounding* means eliciting the judge’s task-relevant knowledge; it does not imply external verification. MERG and the associated statistic Δ_K diagnose how agreement changes under this intervention, not whether the resulting scores are absolutely correct.

The study separates three empirical scopes. The baseline fully crosses 32 LLMs, three frontier judges, 100 writing tasks, and 11 temperatures, yielding 105,600 evaluations. The controlled Original MERG comparison and rubric ablations cover the two models with the highest baseline scores, DeepSeek-R1 and Qwen3-235B, using the same

prompts and judge versions at $t=0.0$. A separate Universal-rubric sweep spans all 11 temperatures but contains missing outputs and uses a different Gemini configuration; we therefore treat it as exploratory sensitivity evidence.

Our experiments yield three principal findings:

1. **Agreement falls under Original MERG.** At $t=0.0$, mean agreement decreases from 0.615 to 0.224 for DeepSeek-R1 and from 0.673 to 0.254 for Qwen3-235B (Table 3). The exploratory Universal sweep shows the same mean direction across temperature, but its design does not support a controlled causal estimate.
2. **The Rubric Commensurability Problem (§3.3).** Mean agreement is 0.239 with independently generated rubrics and 0.622 with a shared, Claude-generated five-dimension rubric package. Moving to independent rubrics yields a normalized reduction of 62%. This is a joint protocol contrast: the shared condition also fixes descriptions, weights, anchors, indicators, Claude-derived metacognitive reminders, authorship, and scoring procedure.
3. **The Resolution Paradox (§3.2).** Model-level Spearman agreement of at least 0.984 coexists with criterion-level Pearson agreement of 0.72. These are non-equivalent summaries computed with different metrics and sample sets, so their 0.27 difference is descriptive rather than a formal decomposition. Nevertheless, the contrast shows that near-identical model rankings need not imply equally consistent fine-grained judgments.

All judges preserve coarse tier separation, but their finer-grained signals are substantially less consistent. This mismatch is particularly relevant when RLAIIF deploys judges at the response level rather than the model level. Agreement should be validated with the evaluation instrument and at the resolution of deployment. Δ_K provides one way to audit that dependence.

2 Methodology

2.1 Experimental Design

Data. We sample 100 prompts from WritingBench (Wu et al., 2025) by stratified sampling, covering 12 to 19 prompts per domain and 94 of the 100 subdomains, balanced between 50 English

The Anatomy of Evaluation Illusion

Sample #501. Investor pitch for a Chinese online-education platform.

Output excerpt. “AI-driven vocational-education leader”; “subsidy budgets grow 20% annually”; “Online Education Service Certification: EDU2023-XXXX.”

Baseline: high scores under the shared rubric

Gemini 9.9: praises the polished formatting and confident projections.

Claude 9.3: emphasizes professional tone and market presentation.

Dimension means: Strategic Positioning 8.7, Professionalism 8.7, Financial Credibility 7.3.

Neither displayed judge questions the placeholder license, unsupported subsidy claim, or absent compliance pathway.

MERG: 6.20-point score spread

Knowledge elicitation: judges articulate the “Double Reduction” policy and online-education licensing requirements.

GPT-5.1 3.7: flags the missing compliance pathway and weak financial modeling.

Claude 4.5 Opus 6.48: identifies unresolved policy risk and conflicting ARPU calculations.

Gemini 2.5 Pro 9.9: continues to reward presentation and strategic viability.

Figure 1: **Evaluation Illusion in Sample #501.** The same response receives markedly different score patterns under the shared baseline rubric and independently generated MERG rubrics. This illustrates sensitivity to the evaluation instrument rather than establishing which scores are more accurate. Full excerpts appear in Appendix D.4.

and 50 Chinese tasks, so the findings rest on near-complete subdomain coverage rather than a narrow slice. WritingBench spans six domains (Literature, Education, Academic, Finance, Politics, Mixed) from creative fiction to technical reports.

Models. We evaluate 32 LLMs spanning three capability tiers:

- **BASE (8):** Qwen2.5-{7B, 32B}, Qwen3-{4B, 8B, 14B, 30B-A3B}, Llama-3.1-8B, Llama-4-Scout (raw pretrained).
- **INSTRUCT (13):** Qwen2.5-{7B, 32B, 72B}, Qwen3-{4B, 8B, 14B, 30B-A3B, 32B, 235B-A22B}, DeepSeek-V3, InternLM3-8B, Llama-3.1-8B, Llama-4-Scout (instruction-tuned).
- **THINKING (11):** DeepSeek-R1-0528, R1-Qwen3-8B, R1-Distill-{Llama-8B, Qwen-7B,

Qwen-32B}, Qwen3-{4B, 8B, 14B, 30B-A3B, 32B, 235B-A22B} (reasoning and distilled reasoning models).

This spread provides a built-in calibration regime. Base models often exhibit conspicuous deficiencies on which judges should converge, whereas Thinking models occupy the high-scoring region where judgments require finer distinctions. Differences in agreement across tiers are therefore informative rather than incidental. For brevity we write *DeepSeek-R1* for DeepSeek-R1-0528 and *Qwen3-235B* for Qwen3-235B-A22B-Thinking throughout.

Judges. The baseline and original MERG conditions use **Claude 4.5 Opus**, **Gemini 2.5 Pro**, and **GPT-5.1**. Drawing the judges from different providers reduces the chance that agreement is an artifact of a single model family or implementation. The baseline uses WritingBench’s checklist-based rubric with 1 to 10 integer scoring. Every one of the 32 models’ outputs (100 prompts each) is scored by all 3 judges at all 11 temperatures ($t \in \{0.0, 0.1, \dots, 1.0\}$), giving a fully crossed design of $32 \times 100 \times 3 \times 11 = 105,600$ evaluation instances. The exploratory Universal-rubric sweep uses the same Claude and GPT versions but a Gemini 2.0 Flash configuration; we therefore treat it as a sensitivity analysis rather than a controlled estimate of the MERG effect.

Metrics. We measure agreement at three granularities:

- *Criterion-level* Pearson r between judge pairs per model $\times$ temperature cell. Cross-temperature analyses use per-criterion scores ($n \approx 500$) for the 30 models with complete criterion-aligned records, excluding Qwen2.5-72B-Instruct and InternLM3-8B-Instruct; the $t=0.0$ rankings use per-prompt means ($n \approx 100$) for all 32 models (Table 4).
- *Cell-level* absolute agreement via ICC(2,1), a two-way random-effects single-measures metric (Shrout and Fleiss, 1979), computed over all 352 cells (32 models $\times$ 11 temperatures) as an $n \approx 100$ subject $\times$ 3 rater design. Unlike r , it penalizes systematic scoring biases.
- *Model-level* Spearman ρ across model means, which captures ranking consistency.

We use the grand mean judge score $\bar{s}$ as an operational proxy for output quality and report paired differences with Cohen’s d . We define the *Knowledge-Grounding Diagnostic*:

$$\Delta_K = \bar{r}_{\text{MERG}} - \bar{r}_{\text{Baseline}} \quad (1)$$

where $\Delta_K < 0$ means that consensus weakens after knowledge grounding; more negative values indicate a larger loss of consensus under the intervention.

2.2 Metacognitive Enhanced Rubric Generation

We design **MERG**, a four-stage framework that prompts LLM judges to move beyond heuristic-first scoring by making task-relevant knowledge explicit before evaluation (Table 1). MERG serves both as a *diagnostic probe* of baseline consensus and as a *prescriptive framework* for more substantive evaluation.

Stage 1: Knowledge Activation. The judge articulates domain-specific knowledge (genre conventions, quality benchmarks, and common pitfalls) before examining any output, anchoring evaluation in explicit expertise rather than implicit heuristics.

Stage 2: Metacognitive Reflection. The judge identifies potential biases it might default to (e.g., favoring confident tone or professional formatting) and articulates mitigation strategies, operationalizing bias awareness (Wang et al., 2024; Wataoka et al., 2024).

Stage 3: Dynamic Rubric Generation. The judge synthesizes activated knowledge into a task-specific rubric $\mathcal{R}$ with unique, deep dimensions. For example, a Gothic horror task generates “Atmospheric Dread Construction” instead of generic “Style” (Table 7).

Stage 4: Calibrated Evaluation. The judge scores each dimension independently with textual evidence, then performs a final bias self-check. The overall score is the weighted mean across dimensions.¹

Ablation Variants. We compare four variants on the two models with the highest baseline scores (DeepSeek-R1 and Qwen3-235B), summarized in Table 2:

¹Full prompt templates are in Appendix A.

Table 1: **Among the compared protocols, MERG uniquely makes judge knowledge explicit.** Existing methods leave domain knowledge implicit or externally specified; MERG prompts the judge to articulate task-relevant knowledge and potential biases before scoring.

Method	Rubric	Knowledge	Grounding
Direct Score	None	Implicit	N/A
G-Eval	Static	Implicit	Template
Prometheus	Static	Implicit	User
WritingBench	Dynamic	Implicit	Query
MERG	Dynamic	Explicit	Metacog.

Table 2: **Four variants compare how rubric sharing changes agreement.** Each variant holds the task and outputs fixed but changes multiple aspects of rubric generation or reuse (§3.3).

Variant	Dimensions	Source
Original	Independent	Each judge
5-Dim Shared	Shared rubric	Claude
Shared Stages	Shared stages 1–3	Claude
Universal	Shared rubric	Precomputed

1. **Original.** Each judge independently executes all four stages, maximizing judge independence.
2. **5-Dim Shared.** Claude generates one task-specific rubric constrained to five dimensions. All three judges share its dimension names, descriptions, weights, scoring anchors, indicators, and Claude-derived metacognitive reminders, then score each dimension independently.
3. **Shared Stages.** Claude’s outputs from Stages 1–3 are reused by Gemini and GPT, which independently execute Stage 4. We report only the Gemini/GPT pair because Claude authored the shared context.
4. **Universal.** Precomputed rubrics are reused identically across judges and temperatures, representing maximal rubric control.

3 Results

3.1 Deconstructing the Shared Illusion

Table 3 presents the controlled comparison at $t=0.0$. Under the original MERG protocol, where each judge independently executes all four stages, mean agreement falls from 0.615 to 0.224 for DeepSeek-R1 and from 0.673 to 0.254 for Qwen3-235B. Both comparisons use the same 100 prompts and the same three judge versions as the baseline.

Figure 2 provides a separate sensitivity analysis using the Universal-rubric variant across all 11

Table 3: **Original MERG lowers agreement at $t=0.0$.** Mean Pearson r across the three judge pairs on the same 100 prompts. Each judge independently executes all four MERG stages, using the same judge versions as the baseline.

Model	Baseline	Original	Δ_K	Rel.
DeepSeek-R1	.615	.224	−.391	−63.6%
Qwen3-235B	.673	.254	−.419	−62.3%

temperatures. Averaged over the three judge pairs at each temperature, Universal-rubric agreement is lower than the baseline for both models at every temperature. Across all pair–temperature combinations, the mean changes from 0.649 to 0.449 for DeepSeek-R1 and from 0.680 to 0.526 for Qwen3-235B; 61 of 66 individual pair–temperature differences are negative. Across the two models, Universal results are available for 1,637 of 2,200 Claude judgments, 2,169 Gemini judgments, and all 2,200 GPT judgments, so correlations use the available samples for each pair. Because the sweep also uses a different Gemini configuration, it supports the persistence of the mean gap but does not provide a controlled estimate of the original MERG effect.

Interpretation. The controlled $t=0.0$ comparison is consistent with independently generated, knowledge-grounded rubrics surfacing disagreements that a shared baseline instrument obscures. The Universal-rubric sweep suggests that a lower mean agreement can persist across decoding temperatures, but its missingness and judge-version difference require a descriptive interpretation.

Two additional observations provide context. **(1) Domain heterogeneity** (§3.4): the Universal-rubric shift varies by both domain and model rather than following a uniform pattern. **(2) Score–agreement gradient:** across 32 models at $t=0.0$, mean judged score and agreement are negatively correlated ($\rho = -0.513$, $p = 0.003$, $n = 32$; Table 4, full list in Table 8). Agreement decreases from Base (0.81) to Instruct (0.77) and Thinking (0.76), consistent with obvious failures driving convergence on outputs from lower-scoring models while outputs from higher-scoring models admit greater evaluative variation.

3.2 The Resolution Paradox

Model-Level vs. Criterion-Level Agreement.

Model-level rankings are nearly identical across judges at $t=0.0$ (Claude/Gemini $\rho = 0.992$, Claude/GPT 0.992, Gemini/GPT 0.984; all $p <$

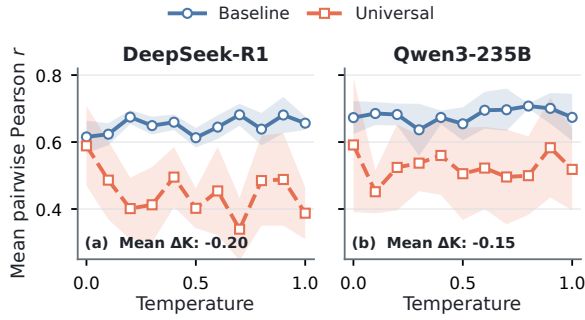


Figure 2: **Universal-rubric agreement remains below baseline across temperature.** Lines show the mean Pearson r over three judge pairs; bands show pairwise standard deviation, not confidence intervals. Correlations use pairwise available samples, so this sweep is descriptive rather than a controlled MERG estimate.

10^{-23}). Across all 32 models and 11 temperatures, the grand mean is $\bar{\rho} = 0.986$ with $SD = 0.006$. At the criterion level, however, Pearson agreement averages only $\bar{r} = 0.72$ across the 30 models with complete criterion-aligned records and 11 temperatures (Appendix C). The descriptive difference of 0.27 is the Resolution Paradox. Judge pairs also differ systematically, with Claude/GPT highest (0.77), followed by Gemini/GPT (0.71) and Claude/Gemini (0.69).

Tier Separation Check. All judges recover the expected ordering: Base $\bar{s} = 3.05$, Instruct $\bar{s} = 5.38$, Thinking $\bar{s} = 5.96$ ($d = 2.42$ over model means, pooled $SD = 1.20$, for Base vs. Thinking; pairwise ordering accuracy 83.9%; Appendix E). The effect is therefore resolution-dependent: coarse tier separation is preserved, while fine-grained within-tier judgments are more fragile. All three judges select the same five highest-scoring models at $t=0.0$, although their internal ordering is not identical, a granularity that may also matter when selecting RLAIIF training pairs.

Three-Level Comparison. ICC(2,1) yields $\overline{ICC} = 0.67$ ($SD = 0.10$), lower than $\bar{r} = 0.72$. The ordering $\rho_{\text{model}} = 0.986 \gg \bar{r}_{\text{criterion}} = 0.72 > \overline{ICC} = 0.67$ provides three increasingly stringent but non-equivalent summaries of agreement; the differences of 0.27 and 0.05 are descriptive rather than a formal decomposition. Both criterion-level and absolute agreement vary with mean judged score: ICC is 0.75 for Base, 0.65 for Instruct, and 0.63 for Thinking at $t=0.0$, while the two highest-scoring models show particularly low absolute agreement (DeepSeek-R1, $ICC = 0.38$;

Table 4: **Mean judged score and agreement move in opposite directions** ($t=0.0$, $\rho = -0.513$, $p = 0.003$, $n = 32$). Left: the 5 highest- and 5 lowest-scoring models by mean judge score $\bar{s}$, with their agreement $\bar{r}$. Right: the 5 highest- and 5 lowest-agreement models. Thinking models dominate the top of the score ranking but exhibit lower agreement on average, whereas Base models dominate the top of the agreement ranking. Full 32-model list in Table 8. B = Base, T = Thinking, Q = Qwen, L = Llama.

By mean score			By agreement		
#	Model	$\bar{s}$	#	Model	$\bar{r}$
1	DeepSeek-R1	7.77	1	Q3-30B-B	.89
2	Q3-235B-T	6.83	2	L4-Scout-B	.88
3	Q3-32B-T	6.72	3	Q3-14B-B	.87
4	Q3-14B-T	6.69	4	Q2.5-7B-B	.87
5	Q3-30B-T	6.62	5	Q2.5-32B-B	.86
28	Q3-4B-B	3.33	28	Q3-235B-T	.67
29	Q2.5-32B-B	2.97	29	Q3-4B-B	.67
30	Q2.5-7B-B	2.20	30	DeepSeek-R1	.62
31	L4-Scout-B	2.04	31	R1-Q3-8B	.60
32	L3.1-8B-B	1.22	32	L3.1-8B-B	.60

Qwen3-235B, $ICC = 0.52$). Criterion-level $\bar{r}$ averaged over temperatures shows the same gradient (Base 0.76 vs. Thinking 0.70). A separate per-prompt analysis of all 32 models at $t=0.0$ likewise yields higher agreement for Base than Thinking models (0.81 vs. 0.76). Model-level summaries can therefore overstate the consistency of the finer-grained signals used in response-level evaluation.

3.3 The Rubric Commensurability Problem

Table 5 shows a sharp difference between independently generated and shared rubrics. Mean agreement is 0.239 when each judge constructs its own rubric and 0.622 when all judges use a Claude-generated, task-specific rubric constrained to five dimensions.

- **Independent rubrics (Original).** When each judge generates its own rubric, agreement is $\bar{r} = 0.224/0.254$ (positive but weak at $n \approx 100$), showing that judges construct substantially different evaluation frameworks.
- **Five-dimension shared rubric.** Sharing complete rubric content raises agreement to 0.543/0.702 ($\Delta r = +0.319/+0.448$). This condition jointly shares dimension names, descriptions, weights, scoring anchors, indicators, and metacognitive context; it does not isolate any one component.

Table 5: **Shared rubric design strongly affects measured agreement.** Pearson r across MERG variants. The 5-Dim Shared condition reuses a complete Claude-generated rubric and Claude-derived metacognitive reminders, not only dimension names; Shared Stages therefore reports only Gemini/GPT. Values are computed from unrounded pairwise correlations.

Method	C/G	C/GPT	G/GPT	Avg
<i>DeepSeek-R1</i>				
Original	.155	.237	.281	.224
5-Dim Shared	.526	.514	.588	.543
Shared Stages	—	—	.606	—
Universal	.678	.668	.420	.589
<i>Qwen3-235B</i>				
Original	.308	.267	.189	.254
5-Dim Shared	.745	.682	.680	.702
Shared Stages	—	—	.649	—
Universal	.747	.719	.308	.591

- **Shared Stages.** When Claude’s Stages 1–3 are reused, the comparable Gemini/GPT correlation is 0.606/0.649, a change of +0.018/−0.031 relative to the five-dimension shared condition.
- **Universal.** Reusing precomputed rubrics yields $\bar{r} = 0.589/0.591$. The asymmetric pairwise pattern— $r = 0.67$ to 0.75 for Claude-involving pairs and 0.31 to 0.42 for Gemini/GPT—shows that tighter rubric control does not guarantee uniformly higher agreement.

Across the two models, moving from the five-dimension shared condition to independently generated rubrics produces a normalized agreement reduction of 62%. Because the conditions differ in rubric content, authorship, dimension constraint, metacognitive reminders, and scoring procedure, this value describes their joint contrast rather than the isolated effect of dimension names. The result nevertheless demonstrates that measured agreement depends strongly on how much of the evaluation instrument judges share (Figure 3).

3.4 Domain- and Model-Dependent Shifts

Domain Analysis. The exploratory Universal-rubric sweep produces different domain profiles for the two models (Figure 4).

- **DeepSeek-R1.** Agreement increases in Academic and Engineering ($\Delta_K = +0.27$), Education (+0.22), and Advertising and Marketing (+0.10), but decreases in Literature and Arts (−0.06), Finance and Business (−0.17), and Politics and Law (−0.20).

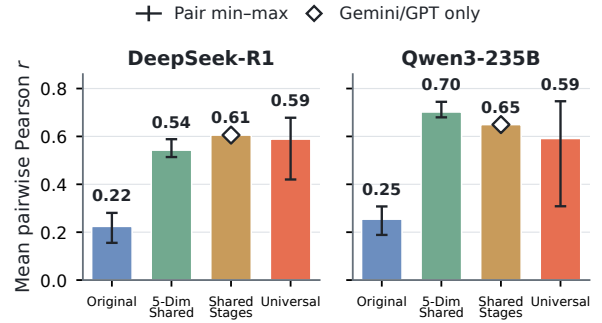


Figure 3: **Agreement depends strongly on rubric sharing.** Bars show mean Pearson r ; whiskers show the judge-pair range. Shared Stages reports only Gemini/GPT because Claude authored the shared context. The contrast changes rubric sharing, dimensionality, and scoring jointly.

Table 6: **The Universal-rubric shift reverses across model-language pairs.** Agreement at $t=0.0$ under the baseline and Universal-rubric protocols. Positive Δ_K denotes increased agreement in the exploratory Universal condition.

Model	Lang	Baseline	Universal	Δ_K
DS-R1	EN	.616	.682	+.07
DS-R1	ZH	.621	.423	−.20
Q3-235B	EN	.800	.608	−.19
Q3-235B	ZH	.543	.638	+.10

- **Qwen3-235B.** Agreement decreases in all six domains, ranging from approximately −0.02 in Education and Academic and Engineering to −0.18 in Advertising and Marketing.

The DeepSeek-R1 profile is consistent with knowledge grounding aligning judges in some codified domains while exposing disagreement elsewhere. Its absence for Qwen3-235B shows that this interpretation is model-dependent rather than a general property of the intervention.

Language Effects. Table 6 shows a model-by-language cross-over in the exploratory Universal-rubric sweep. DeepSeek-R1 gains agreement on English but loses it on Chinese, while Qwen3-235B shows the reverse. The interaction may reflect model-specific language behavior, task composition, rubric variation, missingness, or the Gemini configuration difference and should not be interpreted as a controlled language effect.

Qualitative Evidence. Qualitative analysis of maximum-disagreement cases reveals three recurring patterns: (i) *Logical Verification vs. Formatting* (pitch deck #501; Gemini 9.9 / Claude 6.48 /

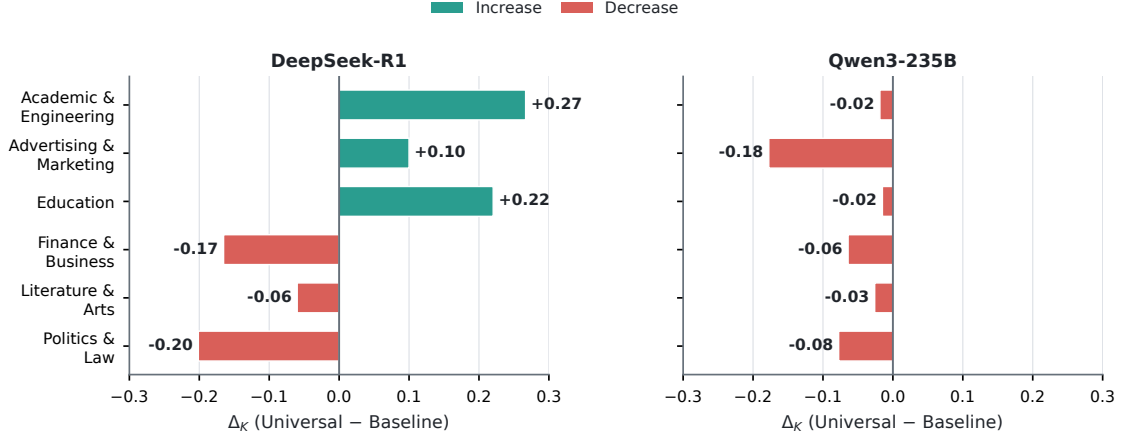


Figure 4: **Universal-rubric shifts vary by domain and model.** At $t=0.0$, $\Delta_K = \bar{r}_{\text{Universal}} - \bar{r}_{\text{Baseline}}$. DeepSeek-R1 increases in three domains and decreases in three, whereas Qwen3-235B decreases in all six. The exploratory sweep includes missing outputs and a different Gemini configuration.

Table 7: **MERG produces more task-specific dimension labels in this example.** Rubric comparison for a Gothic horror writing task (4 of 5 dimensions shown per method). The baseline dimensions are generic, whereas the MERG dimensions refer to conventions of Gothic horror.

WritingBench	MERG
Suspense	Atmospheric Dread Constr.
Coherence	Psychological Ambiguity
Style	The Uncanny Familiar
Grammar	Restraint in Revelation

GPT 3.7), where some judges audit content against domain knowledge while others reward presentation; (ii) *Concept Auditing vs. Keyword Matching* (patent #45; 10.0/6.43/3.2), where two judges detect misuse of zero-knowledge proofs while another matches technical keywords without checking their semantics; and (iii) *Substance vs. Outline* (fiction #739; 9.7/3.0/7.1), where judges differ on whether a structural outline satisfies the requested deliverable. Table 7 contrasts the resulting rubric dimensions; excerpts are in Appendix D.

Heterogeneity and Case Evidence. The Universal-rubric shifts vary across domains, models, and languages rather than following a single pattern. The design limitations of that sweep prevent us from attributing this heterogeneity solely to knowledge grounding. The original MERG case studies nevertheless make the disagreements interpretable: judges identify verifiable failures such as a regulatory contradiction, fabricated financial data, and a misused cryptographic primitive, while differing in

how heavily they weight those failures. Across the 100 DeepSeek-R1 cases analyzed in Appendix D, Gemini assigns the higher score in 90% of its pairwise comparisons with the other judges.

4 Discussion

What Agreement Is Measuring. Taken together, the three results locate the problem in the evaluation pipeline rather than in agreement itself. The Rubric Commensurability Problem shows that sharing an evaluation instrument can substantially change consensus; the Resolution Paradox shows that aggregation can conceal weaker criterion-level signals; and the original MERG comparison shows that agreement can fall when judges independently construct knowledge-grounded rubrics. The exploratory sweep further shows that these shifts vary across models and domains. This complements work on human annotation, where disagreement can carry information about item difficulty (Plank, 2022). For LLM judges, a fluent critique is not evidence of such grounding: detailed justifications can still rest on the same shallow cues.

Implications for Reward Modeling. Prior work documents reward overoptimization (Gao et al., 2023; Rafailov et al., 2024) but leaves open whether the weakness begins in the preference signal. Our results identify two upstream vulnerabilities: a shared rubric can make shallow consensus look reliable, and model-level validation can conceal weaker criterion-level agreement. The Resolution Paradox therefore raises a concern for RLAI, where judges validated by aggregate rankings may be deployed as response-level preference signals.

Auditing those signals with Δ_K and diversifying rubric structures may help reduce this mismatch.

Practical Recommendations.

1. **Audit agreement depth** with Δ_K : a negative value indicates that consensus weakened under the intervention, whereas a positive value indicates increased convergence; either direction requires substantive inspection.
2. **Report rubric provenance**, since agreement induced by a common instrument should not be conflated with independently convergent judgment.
3. **Match evaluation depth to the domain**: report domain- and model-specific effects rather than assuming that knowledge grounding shifts agreement uniformly.
4. **Diversify RLAIIF signals** across rubric structures to reduce potential structural inflation.
5. **Validate at the deployment resolution**: model-level ranking agreement does not establish the reliability of per-sample preference signals.

5 Related Work

LLM Evaluation and Its Limitations. The LLM-as-a-Judge paradigm (Zheng et al., 2023) underpins leaderboards (Chiang et al., 2024), benchmarks (Li et al., 2023, 2024b), chain-of-thought evaluation (Liu et al., 2023), and fine-tuned judges (Zhu et al., 2025); see Li et al. (2024a) for a survey. These works often report aggregate inter-judge agreement as evidence of reliability, but rarely ask whether that agreement is surface-level or substantive. Documented biases include verbosity (Zheng et al., 2023), position (Wang et al., 2024), self-enhancement (Wataoka et al., 2024), and format (Li et al., 2024a, 2025), and JudgeBench (Tan et al., 2025) shows frontier models perform marginally above random on hard evaluation tasks. Closest to us, Wu and Aji (2023) show that judges prefer answers containing factual errors over answers that are merely short or ungrammatical, establishing that style can dominate substance for an *individual* judge. Our contribution is orthogonal and complementary: we ask what happens when judges *agree*. Style-over-substance is a claim about accuracy; Evaluation Illusion concerns the *reliability statistic itself*: inter-judge agreement can

be induced by a shared instrument and may therefore provide less reassurance than it appears to. In the settings studied here, the documented biases are consistent with a shared surface-heuristic repertoire when domain knowledge is not made explicit.

Rubric-Based Evaluation. Structured approaches include Prometheus (Kim et al., 2024a,b), OpenRubrics (Liu et al., 2025), FLASK (Ye et al., 2024), SedarEval (Fan et al., 2025), domain-specific rubrics (Pathak et al., 2025), AutoRubric (Rao and Callison-Burch, 2026), and checklist methods that decompose evaluation into binary items (Lee et al., 2025; Zhang et al., 2026; Li et al., 2024c). Rezaei et al. (2025) elicit rubrics from pairwise comparisons; LitBench (Fein et al., 2025) provides debiased human annotations for creative writing. These methods aim to *improve* evaluation; we instead use rubric variation as a *diagnostic* and show that sharing a rubric can substantially alter measured agreement. Our ablation changes several rubric properties jointly, so it does not isolate which component drives the effect.

Alignment and Metacognition. Gao et al. (2023) established scaling laws for reward overoptimization and Rafailov et al. (2024) extended them to direct alignment algorithms like DPO (Rafailov et al., 2023); Self-Rewarding LMs (Yuan et al., 2024) use the model itself as critic. These works locate divergence *during* optimization; we offer an upstream complement, namely that the reward signal may already be a Shared Illusion beforehand. Self-Refine (Madaan et al., 2023), Reflexion (Shinn et al., 2023), and MENTOR (Shan et al., 2025) apply metacognition to *generation*; MERG applies it to *evaluation*. Finally, Plank (2022) argues that annotation *disagreement* carries signal about item difficulty; we demonstrate the converse, that *agreement* can itself be uninformative.

6 Conclusion

Agreement among LLM judges can be highly reproducible without being knowledge-grounded. Our fully crossed baseline comprises 105,600 evaluations and reveals near-perfect model-level rankings alongside weaker criterion-level and absolute agreement. Separate experiments on two high-scoring models show that agreement also differs sharply between independently generated rubrics and a shared five-dimension rubric package. Under the con-

trolled original MERG comparison, making task knowledge and potential biases explicit before scoring substantially lowers agreement on both intervention models. The exploratory Universal-rubric sweep shows that agreement shifts also vary across models and domains, but its design does not support a causal interpretation of those differences. The practical lesson is not to abandon LLM judges, but to validate them at the resolution and with the evaluation instrument at which they will be used. Consensus should be treated as evidence only after the instrument itself has been examined as a possible source.

Limitations

Our study has several limitations. First, we lack human ground-truth annotations, so we cannot claim that MERG scores are absolutely more accurate, only that agreement weakens under the controlled original MERG comparison; collecting expert judgments on the maximum-disagreement cases is a natural next step. Second, our evaluation covers creative and professional writing via WritingBench, and generalization to exact-answer domains such as code and mathematics warrants separate study. Third, all three judges are proprietary models, so the findings may not transfer to open-source or fine-tuned judges (Zhu et al., 2025). Fourth, the full MERG ablation was limited to the two models with the highest baseline scores (DeepSeek-R1 and Qwen3-235B), so broader evaluation is needed to establish how the effect scales across model families. Fifth, the five-dimension shared condition changes rubric content, authorship, dimensionality, metacognitive reminders, and scoring procedure jointly; it cannot isolate the effect of dimension names. Finally, the Universal-rubric sweep contains incomplete judge outputs and uses a different Gemini configuration from the controlled baseline comparison. Its temperature, domain, and language patterns are therefore exploratory, and Δ_K should be treated as a diagnostic rather than a quantity to be maximized.

Ethics Statement

This work evaluates the reliability of LLM-based evaluation systems and does not involve human subjects or personally identifiable information. All evaluated model outputs are generated from the publicly available WritingBench benchmark. We use proprietary LLM APIs under their standard

terms of service. Our findings highlight potential risks in over-relying on automated evaluation for alignment, which we believe serves the broader goal of responsible AI development. We note that the concept of “Evaluation Illusion” should not be used to dismiss LLM evaluation entirely, but rather to encourage more rigorous evaluation practices. The illustrative example in Figure 1 uses evaluation scores from our experiments applied to a model-generated pitch deck, not a real business plan.

References

- Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E Gonzalez, and Ion Stoica. 2024. Chatbot arena: An open platform for evaluating llms by human preference. *arXiv preprint arXiv:2403.04132*.
- Zhiyuan Fan, Weinong Wang, Xing Wu, and Debing Zhang. 2025. Sedareval: Automated evaluation using self-adaptive rubrics. *arXiv preprint arXiv:2501.15595*.
- Daniel Fein, Sebastian Russo, Violet Xiang, Kabir Jolly, Rafael Rafailov, and Nick Haber. 2025. Litbench: A benchmark and dataset for reliable evaluation of creative writing. *arXiv preprint arXiv:2507.00769*.
- Leo Gao, John Schulman, and Jacob Hilton. 2023. Scaling laws for reward model overoptimization. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 10835–10866. PMLR.
- Daniel Kahneman. 2011. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, New York.
- Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. 2024a. Prometheus: Inducing fine-grained evaluation capability in language models. In *The Twelfth International Conference on Learning Representations*.
- Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024b. Prometheus 2: An open source language model specialized in evaluating other language models. *arXiv preprint arXiv:2405.01535*.
- Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. 2023. RLAIIF vs. RLHF: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*.

- Yukyung Lee, JoongHoon Kim, Jaehee Kim, Hyowon Cho, Jaewook Kang, Pilsung Kang, and Najoung Kim. 2025. [Checkeval: A reliable llm-as-a-judge framework for evaluating text generation using checklists](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*.
- Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhat-tacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. 2024a. From generation to judgment: Opportunities and challenges of llm-as-a-judge. *arXiv preprint arXiv:2411.16594*.
- Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E Gonzalez, and Ion Stoica. 2024b. From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline. *arXiv preprint arXiv:2406.11939*.
- Weiyuan Li, Xintao Wang, Siyu Yuan, Rui Xu, Jiangjie Chen, Qingqing Dong, Yanghua Xiao, and Deqing Yang. 2025. Curse of knowledge: When complex evaluation context benefits yet biases llm judges. *arXiv preprint arXiv:2509.03419*.
- Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. AlpacaEval: An automatic evaluator of instruction-following models. https://github.com/tatsu-lab/alpaca_eval.
- Zongjie Li, Chaozheng Wang, Pingchuan Ma, Daoyuan Wu, Shuai Wang, Cuiyun Gao, and Yang Liu. 2024c. [Split and merge: Aligning position biases in LLM-based evaluators](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*.
- Tianci Liu, Ran Xu, Tony Yu, Ilgee Hong, Carl Yang, Tuo Zhao, and Haoyu Wang. 2025. Openrubrics: Towards scalable synthetic rubric generation for reward modeling and llm alignment. *arXiv preprint arXiv:2510.07743*.
- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522.
- Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegrefe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36.
- Aditya Pathak, Rachit Gandhi, Vaibhav Uttam, Arnav Ramamoorthy, Pratyush Ghosh, Aaryan Raj Jindal, Shreyash Verma, Aditya Mittal, Aashna Ased, Chirag Khatri, Yashwanth Nakka, Devansh, Jagat Sesh Challa, and Dhruv Kumar. 2025. Rubric is all you need: Enhancing LLM-based code evaluation with question-specific rubrics. *arXiv preprint arXiv:2503.23989*.
- Barbara Plank. 2022. 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*, pages 10671–10682.
- Rafael Rafailov, Yaswanth Chittipetu, Ryan Park, Harshit Sikchi, Joey Hejna, W. Bradley Knox, Chelsea Finn, and Scott Niekum. 2024. Scaling laws for reward model overoptimization in direct alignment algorithms. In *Advances in Neural Information Processing Systems*, volume 37.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pages 53728–53741.
- Delip Rao and Chris Callison-Burch. 2026. Autorubric: A unifying framework for rubric-based LLM evaluation on non-verifiable tasks. *arXiv preprint arXiv:2603.00077*.
- MohammadHossein Rezaei, Robert Vacareanu, Zihao Wang, Clinton Wang, Bing Liu, Yunzhong He, and Afra Feyza Akyürek. 2025. Online rubrics elicitation from pairwise comparisons. *arXiv preprint arXiv:2510.07284*.
- Liang Shan, Kaicheng Shen, Wen Wu, Zhenyu Ying, Chaochao Lu, Yan Teng, Jingqi Huang, Qingshan Liu, Guangze Ye, Guoqing Wang, Jie Zhou, and Liang He. 2025. Mentor: A metacognition-driven self-evolution framework for uncovering and mitigating implicit domain risks in llms. *arXiv preprint arXiv:2511.07107*.
- Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36.
- Patrick E Shrout and Joseph L Fleiss. 1979. Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2):420–428.
- Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Y. Tang, Alejandro Cuadron, Chenguang Wang, Raluca Ada Popa, and Ion Stoica. 2025. Judgebench: A benchmark for evaluating llm-based judges. In *The Thirteenth International Conference on Learning Representations*.
- Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024. [Large language models are not fair evaluators](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*.

Koki Wataoka, Tsubasa Takahashi, and Ryokan Ri. 2024. Self-preference bias in llm-as-a-judge. *arXiv preprint arXiv:2410.21819*.

Minghao Wu and Alham Fikri Aji. 2023. Style over substance: Evaluation biases for large language models. *arXiv preprint arXiv:2307.03025*.

Yuning Wu, Jiahao Mei, Ming Yan, Chenliang Li, Shaopeng Lai, Yuran Ren, Zijia Wang, Ji Zhang, Mengyue Wu, Qin Jin, and Fei Huang. 2025. Writingbench: A comprehensive benchmark for generative writing. *arXiv preprint arXiv:2503.05244*.

Seonghyeon Ye, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, Seungone Kim, Yongrae Jo, James Thorne, Juho Kim, and Minjoon Seo. 2024. Flask: Fine-grained language model evaluation based on alignment skill sets. In *International Conference on Learning Representations*.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. 2024. Self-rewarding language models. *arXiv preprint arXiv:2401.10020*.

Zhen Zhang, Kaiqiang Song, Xun Wang, Yebowen Hu, Weixiang Yan, Chenyang Zhao, Henry Peng Zou, Haoyun Deng, Sathish Reddy Indurthi, Shujian Liu, Simin Ma, Xiaoyang Wang, Xin Eric Wang, and Song Wang. 2026. CM2: Reinforcement learning with checklist rewards for multi-turn and multi-step agentic tool use. *arXiv preprint arXiv:2602.12268*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhaghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36.

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

A MERG Prompt Templates

We provide the core prompt templates for each MERG stage. Variables in `{braces}` are filled at runtime.

Stage 1: Knowledge Activation.

You are an expert judge. Before evaluating, activate domain knowledge.

Task Type: `{task_type}`
Query: `{query}`

Activate knowledge in:

1. Task Type Identification
2. Genre-Specific Quality Standards
3. Domain Knowledge Requirements
4. Evaluation Challenges
5. Quality Indicators (good/poor)

Return as structured JSON.

Stage 2: Metacognitive Reflection.

Reflect on your evaluation process:

1. Biases: length, format, familiarity, anchoring, halo
2. Blind Spots: possible omissions?
3. Attention: what needs focus?
4. Mitigation for each bias

Stage 3: Dynamic Rubric Generation.

Using knowledge and bias awareness, generate a task-specific rubric:

- 5 dimensions (NOT generic)
- Each: name, weight, anchors (2=poor, 5=ok, 8=strong, 10=exceptional)
- Each with concrete criteria

Stage 4: Calibrated Evaluation.

Score each dimension independently:

1. Cite specific textual evidence
2. Apply anchors from Stage 3
3. Bias check: did heuristics influence scoring?
4. Compute weighted mean score

B Full Model Rankings

The ranking reveals a clear tier-level pattern. All 8 Base models fall between ranks 20 and 32, occupying 8 of the 13 lowest positions, and their mean agreement ($\bar{r} = 0.81$) exceeds that of Thinking models ($\bar{r} = 0.76$). Individual models depart from this trend: Llama-3.1-8B-Base has the lowest agreement in the table ($r = 0.595$) despite ranking last by score.

C Criterion-Level Agreement

Figure 5 summarizes criterion-level agreement for the 30 models with complete criterion-aligned records at 11 temperature settings ($t \in \{0.0, 0.1, \dots, 1.0\}$). Four patterns are visible:

Temperature Invariance. Across the 30 models, criterion-level Pearson r changes little as temperature varies. This argues against the simple explanation that disagreement arises mainly from stochastic variation in judge outputs; even at $t=0.0$, where sampling variability is minimized, agreement rarely exceeds $r = 0.80$.

Pair-Specific Hierarchy. A consistent hierarchy emerges across models, with Claude/GPT agreement generally highest ($\bar{r} \approx 0.77$), followed by Gemini/GPT ($\bar{r} \approx 0.71$) and Claude/Gemini ($\bar{r} \approx 0.69$). This pattern is stable across Base, Instruct, and Thinking tiers. Gemini also shows the largest average score offset in the case-study sample, although these statistics do not identify its cause.

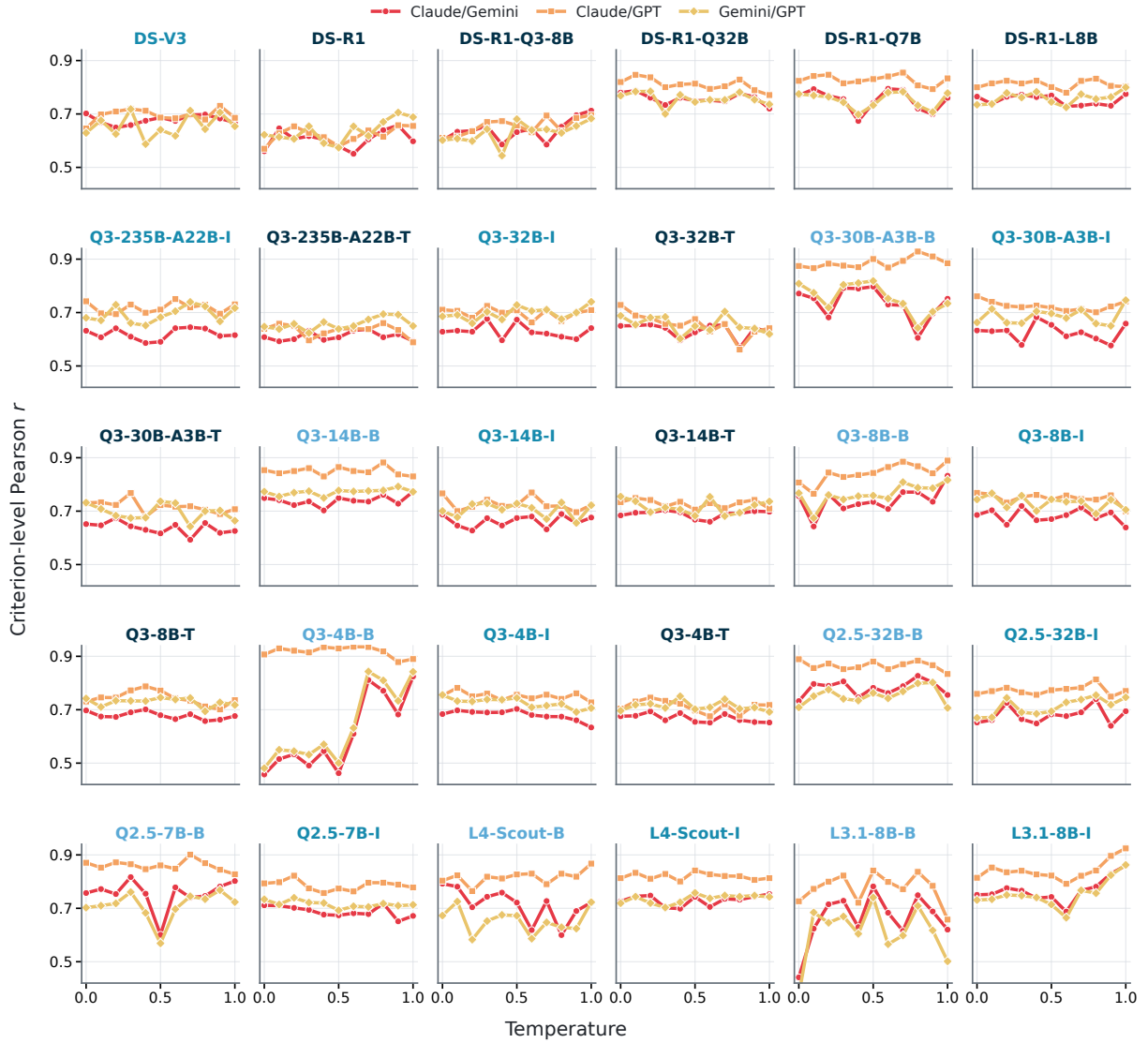


Figure 5: **Criterion-level agreement across 30 models.** Each panel shows three judge-pair correlations over 11 temperatures; title color denotes the model tier (Base, Instruct, or Thinking). Agreement is broadly stable across temperature, while its level varies across models and judge pairs.

Model-Type Effects. Averaged over temperatures, Base models (light blue titles) show *higher* criterion-level agreement ($\bar{r} \approx 0.76$) than Thinking models (dark shading, $\bar{r} \approx 0.70$). This inverse relationship between mean judged score and agreement supports the Resolution Paradox. Outputs from lower-scoring models tend to contain readily identifiable flaws that judges converge on, whereas outputs from higher-scoring models leave more room for differences in evaluative emphasis.

Per-Prompt Cross-Check. A separate analysis of per-prompt mean scores for all 32 models at $t=0.0$ shows the same broad pattern (Base $\bar{r} = 0.81$ vs. Thinking $\bar{r} = 0.76$). The lowest values in this analysis are Llama-3.1-8B-Base ($r = 0.595$)

and DeepSeek-R1-Qwen3-8B ($r = 0.603$). The two highest-scoring Thinking models also fall below the Thinking-tier mean (DeepSeek-R1-0528, 0.615; Qwen3-235B, 0.673), although they are not the two lowest-agreement models overall. Several low-scoring Base models, including Llama-4-Scout-Base and Qwen2.5-7B-Base, show agreement near 0.88.

D Case Studies Under MERG

This section presents representative case studies illustrating how knowledge injection reshapes judge behavior. All cases are drawn from the MERG evaluation of DeepSeek-R1-0528-Thinking and Qwen3-235B-A22B-Thinking, deliberately selected to span the full agreement spectrum, from

Table 8: **Complete 32-model rankings at $t=0.0$.** Mean judged score and agreement are negatively correlated ($\rho = -0.513$, $p = 0.003$). Agreement is lower on average for the Thinking tier than for the Base tier, consistent with the cross-model trend.

Rank	Model	Score	Agr. (r)
1	DeepSeek-R1-0528-Thinking	7.77	.615
2	Qwen3-235B-Thinking	6.83	.673
3	Qwen3-32B-Thinking	6.72	.776
4	Qwen3-14B-Thinking	6.69	.794
5	Qwen3-30B-Thinking	6.62	.780
6	Qwen3-235B-Instruct	6.50	.764
7	Qwen3-32B-Instruct	6.39	.762
8	DeepSeek-V3-0324	6.37	.707
9	Qwen3-8B-Thinking	6.27	.796
10	DeepSeek-R1-Qwen3-8B	6.08	.603
11	Qwen3-30B-Instruct	6.04	.778
12	Qwen3-14B-Instruct	6.03	.779
13	Qwen3-4B-Thinking	5.82	.743
14	Qwen3-8B-Instruct	5.70	.798
15	Qwen2.5-72B-Instruct	5.38	.714
16	Qwen3-4B-Instruct	5.33	.820
17	InternLM3-8B-Instruct	5.09	.772
18	DeepSeek-R1-Distill-Qwen-32B	4.94	.855
19	Qwen2.5-32B-Instruct	4.66	.721
20	Qwen3-14B-Base	4.48	.874
21	Qwen2.5-7B-Instruct	4.42	.803
22	Llama-4-Scout-Instruct	4.41	.812
23	Qwen3-8B-Base	4.26	.830
24	DeepSeek-R1-Distill-Llama-8B	4.22	.845
25	Qwen3-30B-Base	3.91	.894
26	DeepSeek-R1-Distill-Qwen-7B	3.63	.860
27	Llama-3.1-8B-Instruct	3.61	.839
28	Qwen3-4B-Base	3.33	.669
29	Qwen2.5-32B-Base	2.97	.863
30	Qwen2.5-7B-Base	2.20	.873
31	Llama-4-Scout-Base	2.04	.880
32	Llama-3.1-8B-Base	1.22	.595

near-perfect consensus to extreme divergence. Because many WritingBench tasks are in Chinese, some judges produced Chinese-language reasoning; we present these verbatim and provide English translations (marked “translated from Chinese original”).

D.1 Case A: Agreement on Quality

Task. Commercial copywriting for a high-end redwood furniture product detail page.

MERG Scores. Claude: 8.28, Gemini: 8.35, GPT: 8.40 (**spread = 0.12**).

Analysis. After knowledge activation, all three judges identified overlapping substantive strengths while surfacing complementary weaknesses.

- **Claude:** “This copy achieves strong performance across all dimensions, with particular excellence in sensory persuasion and material expertise. The

brand narrative integration is effective though incomplete (missing sub-brand differentiation).”

- **Gemini:** “Exceptional in its literary and narrative craft but flawed in its functional execution. The scores for ‘Persuasive Brand Narrative’ (9.5) and ‘Structural Craftsmanship’ (9.5) are near-perfect [...] The lowest score, ‘Creative Problem-Solving’ (4.0), is a direct consequence of the writer’s failure to resolve the central mahogany/redwood ambiguity.”
- **GPT:** “The most impressive aspect is the detailed, credible depiction of materials and craftsmanship, which is critical for high-end redwood furniture [...] Brand voice is also robust: it consistently communicates heritage, mastery, and exclusivity while avoiding crass sales language.”

Baseline Comparison. At $t=0.0$, baseline judges rewarded Brand Integration and Persuasive Language (both 9.3) while assigning a lower Technical Accuracy score (4.7). Under MERG, all three judges placed greater emphasis on material accuracy and cited concrete evidence for their assessments, yet still converged on the final score. This case illustrates agreement grounded in overlapping substantive evidence rather than generic praise.

D.2 Case B: Agreement on Failure

Task. Financial analysis report on a major gaming company’s overseas market performance.

MERG Scores. Claude: 4.53, Gemini: 4.60, GPT: 4.68 (**spread = 0.15**).

Analysis. All three judges independently identified systematic data fabrication as a critical failure.

- **Claude:** “This response represents a significant failure in financial analysis integrity despite its professional appearance. While it accurately cites verifiable figures from the source document (Q4 revenue, growth rates, mobile proportions, 蛋仔派对 metrics), it extensively fabricates overseas data that the source explicitly does not contain.”
- **Gemini:** “The fatal issue is the systemic data fabrication, which renders the analysis untrustworthy. This rightly results in a very low score (3.0) in the two most heavily weighted dimensions: ‘Accuracy & Fidelity’ (50%) and ‘Analytical Integrity’ (25%).”

- **GPT** (translated from Chinese original): “在任务完成度、结构清晰度和语言表达方面表现明显高于平均水准，完全可以达到“Strong”级别 [...] 然而根据本次 rubric 的硬约束，“财务数据准确性”与“分析深度与不确定性管理”是关键维度：文章在这些方面存在系统问题——大量具体海外营收占比、区域结构，以及单款游戏海外ARPPU/DAU/营收贡献的数字是捏造的，且未做充分的推断标注或来源说明。” (*Performance in task completion, structural clarity, and language expression is clearly above average, fully reaching “Strong” level [...] However, per the rubric’s hard constraints, “Financial Data Accuracy” and “Analytical Depth & Uncertainty Management” are key dimensions: the article has systematic problems: detailed overseas revenue shares, regional structures, and per-game ARPPU/DAU/revenue figures are fabricated without adequate inference labeling or source attribution.*)

Interpretation. This case demonstrates that MERG enables judges to converge on *substantive* failures. In the baseline evaluation (Financial Data Accuracy: 5.0), the fabrication issue was partially masked by the report’s professional formatting. Knowledge activation prompted all three judges to scrutinize the factual support for the report, leading them independently to the same conclusion.

D.3 Case C: Knowledge-Driven Disagreement

Task. Patent disclosure document (专利交底书) based on an academic paper.

MERG Scores. Claude: 6.43, Gemini: 10.0, GPT: 3.2 (spread = 6.80).

Analysis. This extreme disagreement reveals fundamentally different evaluation philosophies activated by domain knowledge.

- **Gemini (10.0):** Evaluated the document purely on *format and rhetorical execution*: “For ‘Technical Depth’, the most heavily weighted dimension, the response provided a wealth of invented, credible, and specific details that perfectly match the ‘Exceptional’ anchor [...] its ‘Structural & Linguistic Compliance’ was perfect.” Gemini’s knowledge activation focused on patent document *structure* rather than *substance*.
- **Claude (6.43):** Balanced format against content: “A document that successfully follows patent dis-

closure conventions and makes reasonable attempts at innovation extraction, but is undermined by significant technical accuracy issues [...] primarily due to the critical misrepresentation of ZKP as an encryption mechanism rather than a proof system.”

- **GPT (3.2)** (translated from Chinese original): “技术正确性受到 ZKP 概念误用和部分随意参数设定的明显拖累，难以给高分 [...] 在专利文体和结构方面则表现较好 [...] 总体评价略高于“中等水平”，视为质量中等偏上但存在明显瑕疵的初稿，给出约 3.2 左右的综合分。” (*Technical correctness is significantly dragged down by ZKP concept misuse and arbitrary parameter settings, making a high score difficult [...] patent style and structure perform well [...] Overall slightly above “medium,” considered a draft of above-average quality but with obvious flaws, yielding a composite score of approximately 3.2.*)

Baseline Comparison. At $t=0.0$, baseline scores showed high agreement (Patent Structure: 8.0, Content Extraction: 7.7). MERG exposed a conflict that the baseline consensus masked: *should a patent disclosure be evaluated primarily on its formal completeness or its technical validity?* The judges reached sharply different conclusions after being prompted to activate task-relevant knowledge.

D.4 Case D: The Formatting Trap

Task. Investor pitch for an online-education platform in China (post-2021).

MERG Scores. Claude: 6.48, Gemini: 9.9, GPT: 3.7 (spread = 6.20).

Analysis. This case (also illustrated in Figure 1) exemplifies the core Evaluation Illusion phenomenon:

- **Gemini (9.9):** “Demonstrates a masterful understanding of the core task: creating a persuasive investment pitch within a complex regulatory environment. It scores perfectly on ‘Strategic Viability’ by turning policy constraints into a competitive advantage [...] This is a benchmark example of an ‘Exceptional’ response.” Gemini activated knowledge about pitch deck conventions but did not fully engage with China’s 2021 “Double Reduction” policy implications.

- **Claude (6.48)** (translated from Chinese original): “内容层面存在几个关键缺陷: 1) 完全忽略双减政策, 这是评估中国EdTech投资的必要背景, 文档虽将公司定位为职业教育但未解释其政策安全性; 2) 多处使用占位符(专利号、省份名称)而非真实数据, 表明这更像是模板而非真实案例; 3) 财务预测存在内部矛盾(ARPU计算值与声称值不符); 4) 竞争分析缺乏客观性和可验证性。” (*Several key flaws at the content level: 1) completely ignores the Double Reduction policy, a necessary context for evaluating Chinese EdTech investment. Though the document positions the company in vocational education, it does not explain its policy safety; 2) multiple placeholders (patent numbers, province names) instead of real data suggest this is more template than genuine case; 3) financial projections contain internal contradictions (ARPU calculations don't match claimed values); 4) competitive analysis lacks objectivity and verifiability.*)
- **GPT (3.7)** (translated from Chinese original): “项目描述为中国职业教育平台, 却完全不回应“双减”等关键监管, 对在线办学资质等合规路径无交代 [...] 财务部分虽然给出两年预测, 却没有按业务线拆分收入, 用户数、ARPU、企业客户之间逻辑混乱, 单位经济缺席, 属于典型“堆数字而不建模型” [...] 材料总体更接近“故事化宣传稿”, 不足以支撑严肃投资判断, 因此将最终综合评价压在4分以下, 定为3.7, 以反映结构亮点与内容不足之间的不平衡。” (*The project is described as a Chinese vocational education platform but completely ignores key regulations like the “Double Reduction” policy, with no explanation of online education licensing compliance [...] Financials give a two-year projection but lack per-business-line revenue splits, with confused logic among user counts, ARPU, and enterprise clients, typical “stacking numbers without building models” [...] Overall closer to a “story-based promotional piece,” insufficient for serious investment decisions, hence the final composite score is set below 4 at 3.7, reflecting the imbalance between structural highlights and content deficiencies.*)

Baseline Comparison. Baseline judges gave uniformly high scores (Strategic Positioning: 8.7, Content & Professionalism: 8.7, Financial Credibility: 7.3). The baseline consensus appears largely driven by surface-level professionalism. Under

MERG, Claude and GPT applied regulatory knowledge while Gemini continued to prioritize presentation, revealing large differences in how the judges weight domain violations.

D.5 Case E: Cross-Model Consistency

Task. Professional film criticism essay comparing two Chinese TV crime dramas.

MERG Scores. Claude: 6.10, Gemini: 5.90, GPT: 5.80 (spread = 0.30).

Analysis. All judges converged on a “competent but flawed” assessment, independently identifying the same critical error:

- **Claude** (translated from Chinese original): “文章在结构组织和叙事分析上表现尚可, 展现了一定的专业素养和比较视野。但存在几个关键问题: 1) 将刘华强误写为‘李天柱’是重大事实错误, 严重影响专业可信度; 2) 分析缺乏具体场景支撑, 多为概括性描述, 华丽修辞背后实质内容有限; 3) 篇幅未达要求, 两剧分析失衡; 4) 结论缺乏更深层的类型片发展洞见。综合来看, 这是一篇中等水平的影视评论, 有基本框架但深度不足, 有专业意识但执行有缺陷。” (*The essay performs adequately in structural organization and narrative analysis, demonstrating a degree of professional awareness and comparative perspective. However, several key issues exist: 1) misidentifying Liu Huaqiang as “Li Tianzhu” is a major factual error that severely undermines professional credibility; 2) analysis lacks specific scene support, mostly consisting of generalizations with limited substance behind ornate rhetoric; 3) length falls short of requirements with imbalanced coverage of the two dramas; 4) conclusions lack deeper insights into crime genre development. Overall, a mid-level film critique with basic structure but insufficient depth, with professional awareness but flawed execution.*)
- **Gemini:** “Its mastery of academic structure, sophisticated language, and analytical method is exceptional, earning scores of 9.0 and 9.5 in those dimensions. On the other hand, it fails the most critical litmus test of the assignment, namely basic factual accuracy about a main character [...] making it a high-quality failure.”
- **GPT** (translated from Chinese original): “各维度得分集中在5–6区间, 表明这是一个“合格但未出众”的回应。比较结构清楚, 叙事

与人物分析方向正确 [...] 然而, 文章的专业性更多体现在语言和宏观判断上, 缺乏与之匹配的细证和具体分析: 叙事技法、人物弧光、关系网络、类型突破等论点少有具体场景支撑, 未能体现 rubric 7–8 分或以上所要求的“扎实例证 + 机制拆解”。” (*Scores cluster in the 5 to 6 range, indicating “adequate but not outstanding.” Comparative structure is clear, with correct direction in narrative and character analysis [...] However, the article’s professionalism is more reflected in language and macro-level judgments, lacking matching detailed evidence and specific analysis: arguments on narrative technique, character arcs, relationship networks, and genre innovation lack specific scene support, failing to meet the rubric’s 7 to 8+ requirement of “solid evidence + mechanism analysis.”*)

Interpretation. This case shows that MERG can preserve agreement even in a subjective domain when judges anchor their assessments in verifiable evidence. Although Literature shows lower agreement for DeepSeek-R1 in the exploratory Universal-rubric analysis ($\Delta_K = -0.06$; Figure 4), the judges converge here on factual accuracy and textual support while retaining room for aesthetic differences.

D.6 Case F: Substance vs. Deliverable

Task. Creative fiction integrating a specified set of prompt elements (Sample #739).

MERG Scores. Claude: 3.0, Gemini: 9.7, GPT: 7.1 (spread = 6.70).

Analysis. Unlike Cases C and D, the disagreement here is not about domain facts but about *what artifact was requested*.

- **Gemini (9.7):** awarded near-perfect marks for “exceptional thematic synthesis” and “a masterful integration of all prompt elements,” evaluating the *concept* the response describes.
- **Claude (3.0):** recognized that the model had produced a sophisticated structural *outline* rather than the requested narrative prose: “An outline, however sophisticated, cannot score above 5 to 6 because it fails to deliver the requested deliverable.”
- **GPT (7.1):** acknowledged the conceptual strength but noted the writing remained at a planning stage, landing between the two.

Interpretation. This case isolates a failure mode that rubric-based scoring can miss: a rubric that omits deliverable compliance can assign high scores to a well-structured outline despite its mismatch with the requested artifact. It exemplifies *Substance vs. Outline* and shows that knowledge activation alone does not guarantee convergence when judges disagree on the unit of assessment.

D.7 Summary of Patterns

Quantitative pattern. Across the 100 DeepSeek-R1 samples evaluated with MERG, Gemini assigned the higher score in 90% of its pairwise comparisons with the other judges (mean 9.0 vs. Claude 7.1, GPT 6.9). Mean pairwise absolute differences were Claude/Gemini 2.32, Claude/GPT 1.10, and Gemini/GPT 2.34; 78% of samples showed a spread >2 points, 41% >3 points, and 19% >4 points. Gemini thus shows the largest positive score offset and participates in the two lower-agreement judge pairs (§3.2); these statistics do not by themselves identify the cause.

Qualitative patterns. Across these cases, three patterns emerge.

1. **Agreement after MERG cites more substantive evidence.** In Cases A, B, and E, the judges cite material accuracy, data fabrication, and factual errors rather than only generic praise for professional tone or clear structure.
2. **Disagreement reveals evaluation philosophy.** In Cases C, D, and F, the disagreement is *interpretable*, reflecting different weightings of format vs. substance, structural completeness vs. technical validity, and concept vs. deliverable.
3. **Baseline agreement is often shallow.** Across the examined cases, baseline agreement is generally higher but more closely tied to surface features that judges can assess without domain expertise.

E Score Trajectories Across Temperatures

Figure 6 displays mean-score trajectories for the 30 models with complete criterion-aligned records across the full temperature range, complementing the all-model $t=0.0$ rankings in Table 8.

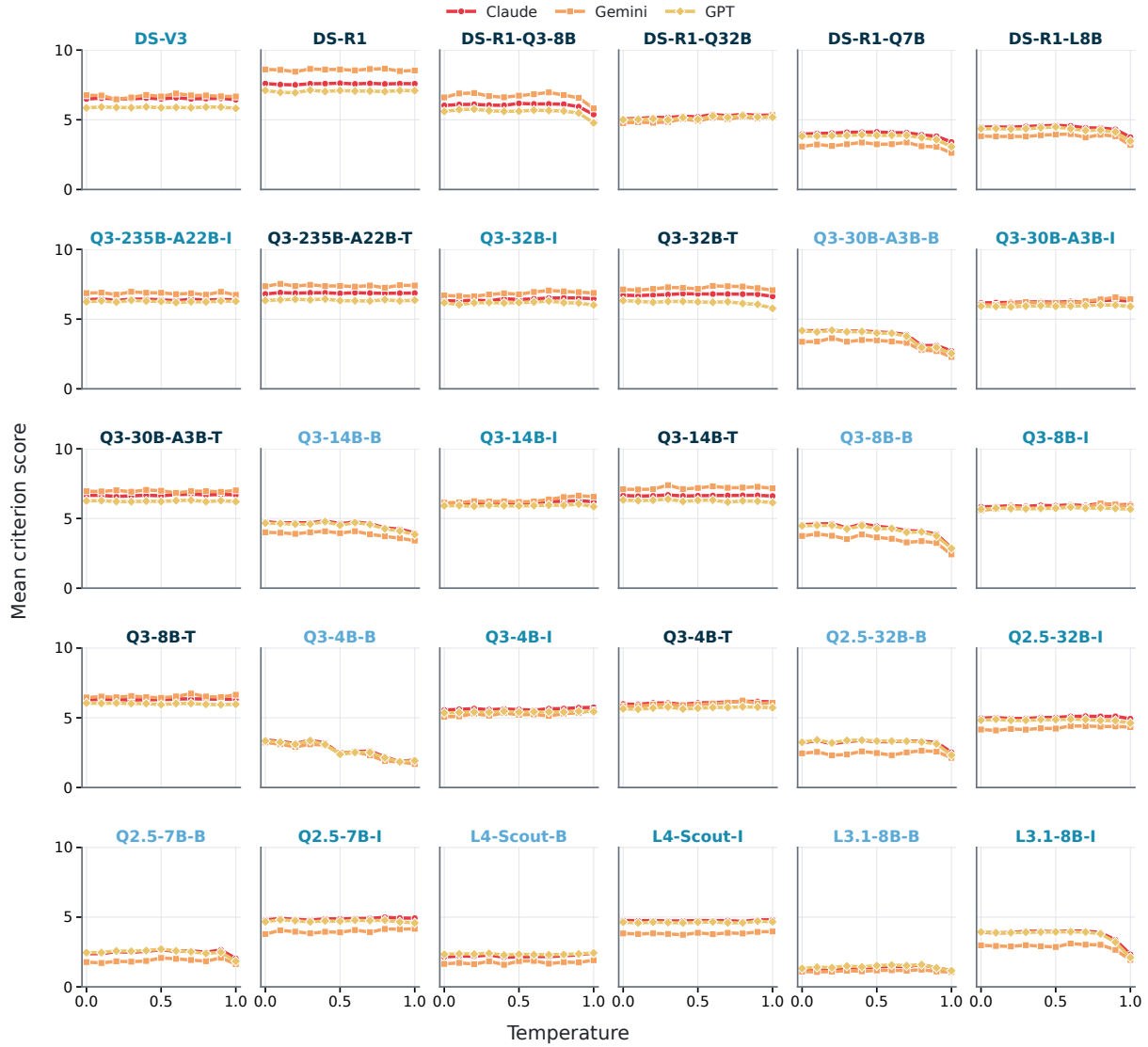


Figure 6: **Coarse tier separation is robust across temperatures.** Mean-score trajectories for the 30 models with complete criterion-aligned records and three judges over $t \in [0.0, 1.0]$. At $t=0.0$, all judges recover the Base < Instruct < Thinking tier ordering (Cohen’s $d = 2.42$ for Base versus Thinking in the full 32-model analysis), even though finer-grained agreement remains weaker.

Three-Tier Score Hierarchy. Model tiers remain well separated. Base models cluster around $\bar{s} = 3.0$ to 5.5 , Instruct models around $\bar{s} = 5.0$ to 7.0 , and Thinking models around $\bar{s} = 5.5$ to 7.5 . All three judges recover this ordering (Cohen’s $d = 2.42$ for Base vs. Thinking), showing that their scores preserve coarse tier separation.

Score Compression among Top Models. Four of the five highest-scoring models cluster between 6.62 and 6.83 , a range of only 0.21 points; DeepSeek-R1 is the exception at 7.77 . This narrow separation is precisely where preference learning must distinguish “good” from “great,” and where small judge-specific effects can invert response-

level rankings.

Judge-Specific Score Patterns. Judge-specific offsets persist across temperatures. Gemini generally assigns the highest means, approximately 0.5 to 1.0 points above Claude and GPT across model tiers, while Claude typically lies between the other two judges. These stable offsets help explain why absolute agreement (ICC) is lower than pairwise correlation even when the trajectories have similar shapes.

Scale and Tuning Effects. Within the Qwen families, larger variants generally receive higher mean scores, but post-training regime matters more than size alone: several smaller Thinking

models outrank larger Instruct and Base variants. DeepSeek-R1-0528-Thinking and Qwen3-235B-A22B-Thinking achieve the two highest mean scores and exhibit agreement below the Thinking-tier mean at $t=0.0$ (0.615 and 0.673, respectively). This pairing reinforces the distinction between tier separation at the model level and weaker agreement at finer granularity.

Implications for RLAIF. Compressed score ranges in the high-scoring tier, judge-specific score offsets, and lower criterion-level agreement could make an RLAIF system more sensitive to judge-specific scoring patterns than to differences captured consistently across judges. This possibility motivates the response-level validation discussed in §4.