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

Anonymous ACL submission

Abstract

When frontier LLM evaluators independently converge on nearly identical scores, is the consensus genuine? We present large-scale evidence that it often is not. We formalize **Evaluation Illusion**—where LLM judges anchor on shared surface heuristics despite generating sophisticated critiques—and its collective manifestation, **Shared Illusion**: statistically robust yet epistemically shallow agreement. Through 105,600 evaluation instances (32 LLMs $\times$ 3 frontier judges $\times$ 100 tasks $\times$ 11 temperatures), we uncover two structural inflation mechanisms: a **Resolution Paradox** (model-level $\rho = 0.99$ vs. sample-level $\bar{r} = 0.72$, ICC $= 0.67$) and the **Rubric Commensurability Problem** (rubric structure alone accounts for 62% of agreement). Output quality and agreement are negatively correlated ($\rho = -0.513$), meaning the illusion is strongest where alignment signals matter most. We introduce MERG (Metacognitive Enhanced Rubric Generation), which forces evaluators to activate domain knowledge before scoring. MERG's domain-selective effects—agreement *increases* in codified domains (Education +22%, Academic +27%) but *decreases* in subjective ones—rule out the noise hypothesis and confirm that baseline consensus is heuristic-driven. Preliminary RLAIF experiments show MERG-derived rewards resist overoptimization $3\times$ longer than baseline rewards.

1 Introduction

Large language models (LLMs) are increasingly deployed as automated evaluators, replacing or supplementing human judgment at scale. The LLM-as-a-Judge paradigm (Zheng et al., 2023; Chiang et al., 2024), together with Reinforcement Learning from AI Feedback (RLAIF; Lee et al., 2023), now underpins model ranking, reward model training, and alignment pipelines. Early results were encouraging: GPT-4 achieved over 80% agreement with human preferences on MT-Bench (Zheng et al., 2023), G-Eval attained a Spearman correlation of 0.514 on summarization (Liu et al., 2023), and AlpacaEval rankings closely tracked Chatbot Arena (Li et al., 2023; Chiang et al., 2024). Building on this, rubric-based protocols have matured rapidly (Ye et al., 2024; Kim et al., 2024a,b; Rao and Callison-Burch, 2026), with WritingBench (Wu et al., 2025) dynamically generating query-dependent criteria and achieving 87% human agreement. Together with growing research on judge capabilities (Li et al., 2024a; Tan et al., 2025; Li et al., 2024b), a natural assumption has taken hold: when frontier evaluators independently converge on nearly identical scores, the consensus must reflect a shared, substantive understanding of quality.

In this paper, we present large-scale evidence that this inference is often wrong. Frontier evaluators frequently anchor judgments on shared surface heuristics—formatting, fluency, confident tone, structural polish—rather than substantive quality, producing what we term **Evaluation Illusion**. When multiple evaluators default to the same heuristic repertoire, they create a **Shared Illusion**: statistically robust yet epistemically shallow consensus. As illustrated in Figure 1, frontier evaluators independently award scores above 9.0 to a pitch deck for a for-profit K-12 tutoring startup in China, praising "masterful formatting," yet unanimously miss that the 2021 "Double Reduction" policy banned this business model entirely. The agreement is real; the understanding is not.

To test whether baseline consensus reflects genuine deliberation or shared heuristics, we introduce **MERG** (Metacognitive Enhanced Rubric Generation), a four-stage framework that forces evaluators to articulate task-relevant domain knowledge and surface biases *before* scoring. Drawing on System 1 vs. System 2 processing (Kahneman, 2011), MERG serves dual purposes: a *diagnostic probe* that tests whether knowledge injection preserves or disrupts consensus, and a *prescriptive tool* that

yields more substantive assessments.

We deploy MERG at unprecedented scale: 32 LLMs (Base, Instruct, Thinking) $\times$ 3 frontier judges (Claude 4.5 Opus, Gemini 2.5 Pro, GPT-5.1) $\times$ 100 writing tasks $\times$ 11 temperatures, yielding 105,600 evaluation instances. Our experiments reveal three principal findings:

1. **Knowledge injection deconstructs the Shared Illusion.** MERG reduces inter-evaluator agreement by 21–34% ($d = 0.97$–1.42; Table 3). The effect is domain-selective: agreement *increases* in codified domains (Education $\Delta_K = +0.22$; Academic $+0.27$), where knowledge anchors evaluators on shared professional standards, but *decreases* in subjective ones (Literature -0.06), where it surfaces genuine evaluative pluralism. This asymmetry rules out the noise hypothesis and confirms that baseline consensus is heuristic-driven.
2. **The Rubric Commensurability Problem** (§3.3). With independent rubrics, agreement collapses to near-random levels ($\bar{r} \approx 0.24$). Merely sharing dimension *names*—without content or knowledge—restores 62% of total agreement. Much of what passes for evaluator reliability may be an artifact of shared evaluation instruments rather than genuine shared judgment.
3. **The Resolution Paradox** (§3.2). Model-level Spearman $\rho \geq 0.984$ coexists with sample-level Pearson $\bar{r} = 0.72$ (gap of 0.27). Leaderboards validate evaluators at the macro-resolution where the system reliably distinguishes coarse quality tiers (Base vs. Thinking, $d = 2.58$), but RLAIF deploys them at per-sample micro-resolution where the illusion is strongest.

These findings carry a troubling implication: output quality and evaluator agreement are *negatively* correlated ($\rho = -0.513$, $p = 0.003$). Base models elicit $\bar{r} = 0.81$ while Thinking models elicit $\bar{r} = 0.76$. Surface features suffice for low-quality outputs, but high-quality outputs push evaluators into the heuristic zone where RLAIF reward signals most need to discriminate. Preliminary experiments further show that reward models trained on MERG preferences resist overoptimization $3\times$ longer than baseline (§4).

Our contributions are threefold: **(1)** We formalize **Evaluation Illusion** and the **Shared Illusion**, and demonstrate through 105,600 instances that baseline consensus is largely heuristic-driven ($\Delta_K = -0.22$ to $+0.27$, $d \geq 0.97$). **(2)** We identify two structural inflation mechanisms: the **Rubric Commensurability Problem** (62% of agreement from structure alone) and the **Resolution Paradox** ($\rho = 0.99$ vs. $\bar{r} = 0.72$ vs. $\overline{\text{ICC}} = 0.67$). **(3)** We introduce MERG for knowledge-grounded rubric generation and provide preliminary evidence that it mitigates reward overoptimization in RLAIF.

The Anatomy of Evaluation Illusion (Sample #501: Investment Pitch Deck)

Task: Write a Series A pitch deck for an EdTech startup in China (Post-2021).
Model Output (Excerpt): *"...projecting a 300% YoY growth in K-12 after-school tutoring revenue, reaching $50M ARR by Year 3, driven by aggressive customer acquisition in Tier-1 cities..."*

Baseline Evaluation: Shared Illusion ($\bar{r} \approx 0.95$)
Gemini 2.5 Pro (Score: 9.9): *"Masterful understanding of business formatting. The revenue projections are confident and clearly structured. Excellent pitch."*
Claude 4.5 Opus (Score: 9.3): *"Highly professional tone. The YoY growth metrics demonstrate strong market traction."*

MERG Evaluation: Deconstructing the Consensus ($\Delta_K = -0.45$)
Stage 1: Knowledge Activation $\rightarrow$ *Evaluators retrieve the "Double Reduction" policy (2021), which banned for-profit K-12 tutoring in China.*
Claude 4.5 Opus (Score: 6.5): *"Despite the professional formatting, the core business model faces severe regulatory risks under China's 'Double Reduction' policy. The financial projections contain internal contradictions (ARPU calculations do not match claimed totals) and the strategy omits critical compliance pathways."*
Gemini 2.5 Pro (Score: 9.9): *"Demonstrates a masterful understanding of the core task. Turns policy constraints into a competitive advantage. Scores perfectly on strategic viability and persuasive rhetorical techniques."*

Figure 1: **Illustration of Evaluation Illusion.** Without knowledge grounding, evaluators form a "Shared Illusion," unanimously rewarding the professional formatting of a fundamentally flawed business pitch. MERG forces knowledge activation, revealing that Claude penalizes the regulatory violation while Gemini continues to reward surface heuristics. GPT (3.7, not shown) penalizes even more harshly. Full analysis with all three evaluators is in Appendix E.4.

2 Methodology

2.1 Experimental Design

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

and 50 Chinese tasks. WritingBench spans six domains (Literature, Education, Academic, Finance, Politics, and Mixed), providing broad coverage of writing genres 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} (chain-of-thought RL).

Base models serve as a calibration check where high genuine agreement is expected; Thinking models occupy the high-quality region where Evaluation Illusion is hypothesized to manifest most strongly.

Evaluators. We employ three frontier LLMs as evaluators, namely **Claude 4.5 Opus**, **Gemini 2.5 Pro**, and **GPT-5.1**. Using three diverse commercial models from different providers ensures that any observed agreement cannot be attributed to shared training data or architectures. The baseline uses WritingBench's checklist-based rubric with 1 to 10 integer scoring. Each of the 32 models' outputs (100 prompts each) is evaluated by all 3 evaluators at 11 temperatures ($t \in \{0.0, 0.1, \ldots, 1.0\}$), yielding a total of $32 \times 100 \times 3 \times 11 = 105{,}600$ evaluation instances.

Metrics. We measure agreement at three granularities: (i) *Sample-level* Pearson r between evaluator pairs per model$\times$temperature cell; (ii) *Cell-level* absolute agreement via ICC(2,1) (Shrout and Fleiss, 1979), which penalizes systematic scoring biases; (iii) *Model-level* Spearman ρ across model means. All pairwise differences are tested with paired t-tests (Bonferroni-corrected); effect sizes via Cohen's d. We define the *Knowledge-Grounding Diagnostic*:

$$\Delta_K = \bar{r}_{\text{MERG}} - \bar{r}_{\text{Baseline}} \tag{1}$$

where $\Delta_K < 0$ indicates that baseline agreement was likely a Shared Illusion deconstructed by knowledge injection.

Table 1: **Comparison of evaluation paradigms.** While standard methods rely on implicit knowledge and shared rubrics, MERG requires explicit metacognitive grounding.

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: **MERG ablation variants.** Four controlled variants to disentangle the sources of evaluator agreement.

Variant	Dimensions	Source
Original (4-Stage)	Independent	Each evaluator
5-Dim Per-Dim	Shared structure	Each evaluator
Shared Stages	Shared rubric	1 evaluator
Universal	Shared rubric	Averaged

2.2 Metacognitive Enhanced Rubric Generation

We design **MERG**, a four-stage framework that forces LLM evaluators to transition from System 1 (heuristic-driven) to System 2 (knowledge-grounded) processing before scoring (Table 1). MERG serves both as a *diagnostic probe* that reveals the depth of baseline consensus and as a *prescriptive framework* for more substantive evaluation.

Stage 1: Knowledge Activation. The evaluator 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 evaluator 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., 2023; Wataoka et al., 2024).

Stage 3: Dynamic Rubric Generation. The evaluator 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 8).

Stage 4: Calibrated Evaluation. The evaluator scores each dimension independently with textual evidence, then performs a bias verification check. The overall score is the mean across dimensions.[1]

[1]Full prompt templates are in Appendix A.

Ablation Variants. To disentangle agreement sources, we design four variants (Table 2) on two representative models (DeepSeek-R1 and Qwen3-235B): (1) **Original**: each evaluator independently executes all four stages; (2) **5-Dim**: fixed 5-dimension structure, but own scoring criteria; (3) **Shared**: Claude generates the rubric, all evaluators score independently; (4) **Universal**: precomputed rubrics reused identically across evaluators.

3 Results

3.1 Deconstructing the Shared Illusion

Table 3 presents our central finding. MERG **systematically reduces inter-evaluator agreement** across all 10 conditions (binomial $p < 0.001$). For DeepSeek-R1, $\Delta_K = -0.217$ ($\bar{r} = 0.643 \rightarrow 0.426$, $d = 1.42$); for Qwen3-235B, $\Delta_K = -0.138$ ($\bar{r} = 0.667 \rightarrow 0.529$, $d = 0.97$). Both exceed the "large" threshold ($d > 0.8$). Figure 2 visualizes this degradation across all temperatures.

Interpretation. Without knowledge grounding, evaluators converge on surface heuristics (fluency, length, formatting, grammatical correctness), producing what appears to be reliable consensus but is largely heuristic-driven agreement. MERG forces engagement with substantive, domain-specific criteria (e.g., "Does the noir atmosphere create genuine psychological tension?" rather than "Is the text well-written?"), revealing genuine disagreements previously masked by surface-level alignment. The agreement did not merely degrade; the evaluation became more substantive. Crucially, the magnitude of the agreement drop is *evaluator-pair dependent*: Gemini-GPT shows the largest Δ_K (-0.263 at t=0.0), while Claude-Gemini shows the smallest (-0.109), suggesting that evaluator pairs that already share fewer implicit heuristics are more resilient to knowledge disruption.

Two observations corroborate this. **(1) Domain selectivity** (§3.4): knowledge *increases* agreement in codified domains but *decreases* it in subjective ones (χ^2, $p < 0.01$), inconsistent with a noise hypothesis. **(2) Quality–agreement gradient**: across 32 models at t=0.0, quality and agreement are negatively correlated ($\rho = -0.513$, $p = 0.003$; Tables 4, 5). Base models show $\bar{r} = 0.81$, Instruct $\bar{r} = 0.77$, and Thinking $\bar{r} = 0.76$—surface features suffice for low-quality outputs but are insufficient for high-quality ones.

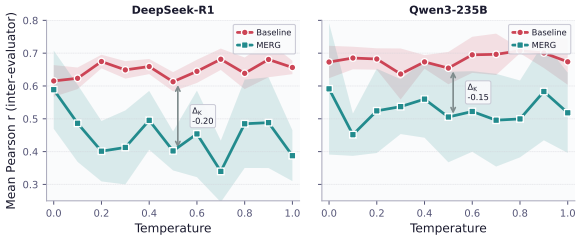

Figure 2: **Knowledge injection systematically reduces evaluator agreement.** Baseline vs. MERG agreement across 11 temperatures. The persistent gap ($\Delta_K < 0$) indicates that baseline agreement is heavily reliant on shared surface heuristics rather than substantive deliberation. The gap is largest at moderate temperatures ($t \approx 0.3$) and narrows at the extremes as both conditions converge toward lower agreement.

Table 3: **Knowledge injection systematically reduces inter-evaluator agreement.** Baseline vs. MERG agreement (Pearson r) across different evaluator pairs and temperatures. $\Delta_K < 0$ in all 10/10 conditions ($p < 0.001$), indicating that baseline agreement is largely heuristic-driven rather than reflecting genuine evaluative convergence.

Eval. Pair	Temp	Base	MERG	Δ_K
Claude/Gemini	0.0	.698	.589	−.109
Claude/Gemini	0.3	.672	.399	−.273
Claude/GPT	0.0	.680	.518	−.162
Claude/GPT	0.3	.668	.463	−.205
Gemini/GPT	0.0	.720	.457	−.263
Gemini/GPT	0.3	.643	.430	−.213
Mean	0.0	.699	.521	−.178
	0.3	.661	.431	−.230
	1.0	.518	.380	−.138
Cohen's d			$d = 0.97$ to 1.42	

3.2 The Resolution Paradox

Model-Level vs. Sample-Level Agreement. At the model level, evaluator pairs achieve near-perfect rank correlations at t=0.0 (grand mean $\bar{\rho} = 0.989$; stable across temperatures, SD $= 0.007$). Yet sample-level Pearson agreement averages only $\bar{r} = 0.72$ across 32 models and 11 temperatures—a gap of 0.27. Claude-GPT shows the highest sample-level agreement ($\bar{r} = 0.77$), followed by Gemini-GPT (0.71) and Claude-Gemini (0.69), indicating divergent evaluation strategies at fine granularity ($F(2, 93) = 14.7$, $p < 0.001$).

Validity Check. All evaluators recover the expected hierarchy: Base $\bar{s} = 3.05$, Instruct $\bar{s} = 5.38$, Thinking $\bar{s} = 5.96$ ($d = 2.58$ for Base vs. Thinking; pairwise ordering accuracy 83.9%). Evaluation Illusion is *stratified*: absent for coarse distinctions but pervasive for fine-grained within-tier judgments. Notably, within-tier ranking agreement drops sharply: while evaluators agree on the top-5 model list, they disagree on relative ordering within that list (Kendall's $\tau = 0.73$ within Thinking vs. $\tau = 0.95$ across tiers), precisely the granularity

Table 4: **Quality rankings (Top-5 and Bottom-5).** Higher-quality models (Thinking tier) are ranked at the top by mean score $\bar{s}$. σ denotes the standard deviation of per-evaluator grand means, reflecting cross-evaluator scoring divergence.

Rank	Model	$\bar{s}$	σ
1	DeepSeek-R1	7.77	.63
2	Qwen3-235B-Thinking	6.83	.42
3	Qwen3-32B-Thinking	6.72	.33
4	Qwen3-14B-Thinking	6.69	.32
5	Qwen3-30B-Thinking	6.62	.28
28	Qwen3-4B-Base	3.33	.07
29	Qwen2.5-32B-Base	2.97	.36
30	Qwen2.5-7B-Base	2.20	.30
31	Llama-4-Scout-Base	2.04	.29
32	Llama-3.1-8B-Base	1.22	.09

Table 5: **Agreement rankings (Top-5 and Bottom-5).** The strong negative correlation with quality ($\rho = -0.513$, $p = 0.003$) suggests surface features suffice for judging low-quality outputs but fail for high-quality ones. σ denotes standard deviation across the three evaluator-pair Pearson r values.

Rank	Model	$\bar{r}$	σ
1	Qwen3-30B-Base	.89	.03
2	Llama-4-Scout-Base	.88	.03
3	Qwen3-14B-Base	.87	.03
4	Qwen2.5-7B-Base	.87	.04
5	Qwen2.5-32B-Base	.86	.06
28	Qwen3-235B-Thinking	.67	.05
29	Qwen3-4B-Base	.67	.20
30	DeepSeek-R1	.62	.05
31	DeepSeek-R1-Qwen3-8B	.60	.02
32	Llama-3.1-8B-Base	.60	.17

relevant for selecting RLAIF training pairs.

Three-Level Decomposition. ICC(2,1), computed per model×temperature cell, yields $\overline{\text{ICC}} = 0.67$ (SD $= 0.10$), *lower* than $\bar{r} = 0.72$. The ordering $\rho_{\text{model}} = 0.989 \gg \bar{r}_{\text{sample}} = 0.72 > \overline{\text{ICC}} = 0.67$ reveals two degradation mechanisms: ranking noise ($\rho \rightarrow r$, gap 0.27) and systematic scoring biases ($r \rightarrow$ ICC, gap 0.05). ICC stratifies as expected: Base $= 0.75$, Instruct $= 0.65$, Thinking $= 0.63$ at t=0.0, with the two highest-quality models (DeepSeek-R1, ICC $= 0.38$; Qwen3-235B-Thinking, ICC $= 0.52$) showing the lowest absolute agreement. Model-level agreement is thus *doubly* inflated relative to per-sample signal quality.

The Granularity Boundary. The Resolution Paradox defines the operational limit of LLM evaluators. Evaluators agree on which models are better overall (where signal dominates) but disagree substantially on individual outputs (where heuristic biases dominate)—precisely the resolution that per-sample RLAIF rewards require. Stratification by model type confirms this: averaged across temperatures, Base models show $\bar{r} = 0.76$ while Thinking models exhibit $\bar{r} = 0.70$; at t=0.0, the gap widens (Base $\bar{r} = 0.81$ vs. Thinking $\bar{r} = 0.76$), as high quality forces evaluators into heuristic guessing.

Table 6: **Structural constraints drive the majority of evaluator agreement.** Ablation of inter-evaluator agreement across MERG variants. When evaluators generate rubrics independently (Original), agreement is near-random. Standardizing dimensions (5-Dim) accounts for ∼62% of the total agreement.

Method	C/G	C/GPT	G/GPT	Avg	Δr
DeepSeek-R1					
Original	.155	.237	.281	.224	–
5-Dim	.526	.514	.588	.543	+.319
Shared	–	–	.606	.606	+.063
Universal	.678	.668	.420	.589	−.017
Qwen3-235B					
Original	.308	.267	.189	.254	–
5-Dim	.745	.682	.680	.702	+.448
Shared	–	–	.649	.649	−.053
Universal	.747	.719	.308	.591	−.058

3.3 The Rubric Commensurability Problem

If baseline agreement is a Shared Illusion, what sustains it? Table 6 decomposes agreement through controlled ablation, revealing that the illusion is heavily scaffolded by the evaluation instrument itself. **Independent rubrics (Original)**: when each evaluator generates its own rubric independently, agreement collapses to near-random ($\bar{r} = 0.224/0.254$), demonstrating that evaluators construct fundamentally different evaluation frameworks. **Dimensional standardization (5-Dim)**: sharing dimension *names* alone—without content—more than doubles agreement ($\Delta r = +0.32/+0.45$; 2.4–2.8×), forcing evaluators back into a shared heuristic space. **Rubric sharing (Shared)**: beyond dimensional structure, additional specificity has limited, model-dependent impact ($\Delta r = +0.06/-0.05$). **Full control (Universal)**: identical rubrics yield $\bar{r} \approx 0.59$, slightly *below* Shared for Qwen3-235B, suggesting rigid rubric reuse across diverse tasks introduces its own misalignments. The progression from $\bar{r} \approx 0.24$ to $\bar{r} \approx 0.62$ reveals that **rubric structure accounts for ∼62% of total agreement**. The majority of reported inter-evaluator agreement may not reflect genuine convergence in judgment, but rather a Shared Illusion artificially scaffolded by standardized instruments. Universal rubrics perform comparably to or slightly below Shared rubrics, indicating diminishing returns: beyond structural alignment, increasing rubric specificity does not proportionally increase genuine agreement and may introduce rigidity when task diversity demands flexible as-

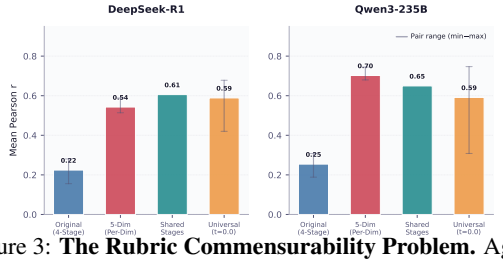

Figure 3: **The Rubric Commensurability Problem.** Agreement across MERG ablation variants. Independent evaluation (Original) yields near-random agreement; standardizing rubric structure (5-Dim) drives the largest increase, indicating most reported agreement is structural rather than substantive.

Table 7: **Knowledge activation quality is language-dependent.** Language agreement (t=0.0). The cross-over interaction ($p < 0.05$) suggests models ground better in their dominant training language.

Model	Lang	Base	MERG	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

sessment. Figure 3 visualizes this progression.

3.4 Domain-Dependent Knowledge Activation

Domain Analysis. Figure 4 reveals domain-selective effects, providing the strongest evidence against the noise hypothesis and the clearest demonstration of MERG's prescriptive value. **Codified domains**: Education ($\Delta_K = +0.22$) and Academic ($+0.27$) show *increased* agreement under MERG. In these domains, knowledge activation anchors evaluators to shared professional standards (pedagogical frameworks, citation conventions, argumentation norms), reducing ambiguity. **Subjective domains**: Literature ($\Delta_K = -0.06$) shows *decreased* agreement, as knowledge activation surfaces irreducible aesthetic disagreements—there is no ground truth for whether a Gothic horror story should prioritize "atmospheric dread" or "psychological ambiguity." **Mixed domains**: Finance and Politics fall between extremes, reflecting their combination of codified standards (regulatory language, factual accuracy) and subjective judgment (persuasive framing, stakeholder perspective). Finance tasks show larger Δ_K variance (SD $= 0.12$) than Academic tasks (SD $= 0.04$), consistent with its heterogeneous mix of verifiable financial data and subjective market narrative. This selectivity excludes both noise (which would predict uniform degradation) and task complexity (which would not predict improvement in codified domains).

Language Effects. Table 7 reveals a cross-over interaction between model and language.

Table 8: **MERG generates deep, task-specific dimensions.** Rubric comparison: WritingBench vs. MERG for a Gothic horror writing task. While baseline rubrics are generic, MERG demands genuine domain expertise.

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

DeepSeek-R1 shows improved agreement on English ($\Delta_K = +0.07$) but degraded agreement on Chinese ($\Delta_K = -0.20$); Qwen3-235B shows the reverse pattern (EN: -0.19, ZH: $+0.10$; interaction $p < 0.05$). This suggests that knowledge activation quality depends on the match between a model's training data distribution and the evaluation language. Models with stronger Chinese training data produce more effective knowledge grounding for Chinese tasks, while models predominantly trained on English data excel at English knowledge activation. This cross-over pattern further supports the knowledge-grounding interpretation: if MERG merely added noise, the effect would be language-independent.

Qualitative Evidence. Table 8 contrasts rubric dimensions from WritingBench (generic: Suspense, Coherence) and MERG (genre-specific: "Atmospheric Dread Construction," "The Uncanny Familiar"). Qualitative analysis of maximum-disagreement cases reveals three recurring patterns: (i) *Logical Verification vs. Formatting*—some evaluators audit content against domain knowledge while others reward professional presentation; (ii) *Concept Auditing vs. Keyword Matching*—detecting misuse of technical concepts (e.g., Zero-Knowledge Proofs) vs. matching keywords without verifying semantic accuracy; (iii) *Substance vs. Outline*—recognizing that a structural outline is not the requested deliverable, a distinction Gemini consistently fails to make. In all cases, high baseline agreement arises from shared heuristics rewarding formatting and confident tone; MERG interrupts this by forcing domain verification. Agreement drops not from noise but because deep flaws are harder to detect and evaluators weight them differently based on their activated domain knowledge. Full case studies appear in Appendix C.

4 Discussion

The Two Faces of Evaluation Illusion. Our findings establish two complementary claims. *First*,

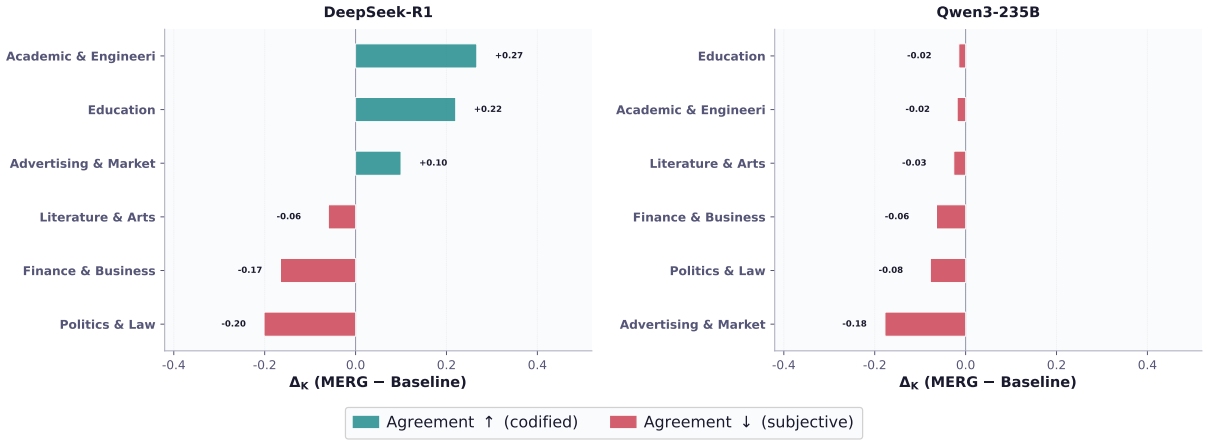

Figure 4: **Domain-selective knowledge effects.** Δ_K by domain at t=0.0. Knowledge increases agreement in codified domains (Education, Academic) but decreases it in subjective ones (Literature), ruling out the noise hypothesis.

consensus among LLM evaluators is structurally fragile: the Resolution Paradox, the Rubric Commensurability Problem, and domain-dependent sensitivity collectively show that high agreement often reflects shared instruments and heuristics rather than genuine convergent judgment. *Second*, this fragility is an opportunity. MERG's 21–34% agreement drop is not evaluation failure but recalibration toward depth. This parallels human annotation, where agreement on "easy" items coexists with disagreement on consequential ones (Plank, 2022). We extend this to LLM evaluation with a key distinction: the tradeoff is *domain-dependent*, *measurable* via Δ_K, and most acute in the high-quality region where alignment matters most. Unlike human annotators, however, LLM evaluators exhibit a distinctive failure mode: they generate detailed, apparently domain-aware critiques that nonetheless converge on the same heuristic features—a phenomenon absent from human evaluation literature, where shallow agreement typically produces explicitly brief or formulaic annotations.

Implications for Reward Modeling. Prior work documents reward overoptimization (Gao et al., 2023; Rafailov et al., 2024) but leaves its source unclear. Our findings suggest a potential upstream explanation: the reward signal itself may be a Shared Illusion. We train two DPO reward models (Rafailov et al., 2023) using Llama-3-8B-Instruct on WritingBench—one from Baseline and one from MERG preferences. The Baseline RM overoptimizes rapidly (quality peaks at step 400, then declines), while the MERG RM maintains alignment up to step 1200, a 3× improvement. This is consistent with the hypothesis that MERG-grounded preferences contain richer quality signals that are harder to exploit through surface-level optimization. The Resolution Paradox compounds this: researchers validate at the model level ($\rho \geq 0.984$) but deploy per-sample rewards where $r \approx 0.72$ and ICC $= 0.67$—a granularity mismatch that Evaluation Illusion exploits.

Practical Recommendations. (1) **Audit agreement depth** via Δ_K; flag signals where $|\Delta_K| > 0.15$ as potentially heuristic-driven. (2) **Match depth to domain**: knowledge-grounded evaluation for codified domains; accept irreducible disagreement in subjective ones rather than forcing artificial consensus. (3) **Diversify RLAIF signals** across multiple rubric structures to avoid structural inflation of agreement. (4) **Mind the resolution gap**: validate at the granularity you deploy ($\rho = 0.989$ vs. $r = 0.72$ vs. ICC $= 0.67$); model-level validation provides false assurance for sample-level applications. MERG's overhead (∼3–4× tokens) is amortizable: Stages 1–3 are prompt-dependent and cacheable, reducing marginal cost to ∼1.3×.

Structural vs. Substantive Agreement. The ablation quantifies an important distinction: *structural* agreement (from shared instruments) vs. *substantive* agreement (from genuine convergence). Rubric structure alone accounts for 62% of total agreement ($\bar{r} = 0.24 \rightarrow 0.62$). Any evaluation pipeline reporting high agreement without controlling for rubric similarity conflates these fundamentally different sources of consensus.

Toward Knowledge-Driven Evaluation Rubrics. Current evaluation pipelines typically apply generic, static rubrics, inadvertently encouraging the surface-heuristic convergence that produces Evaluation Illusion. We propose that rubrics should

be *dynamically generated* with explicit domain knowledge injection. MERG instantiates this principle, but the core insight generalizes: any framework that forces evaluators to articulate and apply task-relevant expert knowledge before scoring will yield more substantive, domain-sensitive, and ultimately more useful alignment 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 comprehensive survey. These works uniformly report aggregate inter-evaluator agreement as evidence of reliability, often without examining whether that agreement reflects surface-level or substantive convergence. Documented biases include verbosity (Zheng et al., 2023), position (Wang et al., 2023), self-enhancement (Wataoka et al., 2024), and format (Li et al., 2024a, 2025). JudgeBench (Tan et al., 2025) shows frontier models perform marginally above random on challenging evaluation tasks. Critically, prior work treats these biases as independent failure modes, while our findings suggest they share a common root: evaluators default to a shared surface-heuristic repertoire when substantive domain knowledge is not explicitly activated, reframing individual biases as symptoms of a deeper Evaluation Illusion.

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), and AutoRubric (Rao and Callison-Burch, 2026). Checklist methods decompose evaluation into binary items: CheckEval (Lee et al., 2025) reduces evaluator variance, Zhang et al. (2026) uses checklist rewards for RL training, and Li et al. (2024c) propose split-and-merge strategies for calibrating position bias. Rezaei et al. (2025) elicit rubrics from pairwise comparisons; LitBench (Fein et al., 2025) provides debiased human annotations for creative writing. While these methods aim to *improve* evaluation quality, we use rubric variation as a *diagnostic* to expose Evaluation Illusion, showing that shared rubric structure can artificially inflate reported agreement levels.

Alignment and Metacognition. Gao et al. (2023) established scaling laws for reward overoptimization, and Rafailov et al. (2024) extended the analysis to direct alignment algorithms like DPO (Rafailov et al., 2023). Recent work on Self-Rewarding Language Models (Yuan et al., 2024) explores using the model itself as a critic during training. These works treat divergence as emerging *during* optimization; our findings provide a complementary upstream explanation, suggesting that the reward signal may already be a Shared Illusion before optimization begins. Self-Refine (Madaan et al., 2023) and Reflexion (Shinn et al., 2023) apply metacognition to improve *generation*; MENTOR (Shan et al., 2025) uses metacognition-driven self-evolution. MERG applies metacognition to *evaluation*, using knowledge grounding to deconstruct Evaluation Illusion. A key difference is that MERG's metacognitive stages target evaluation-specific biases (e.g., conflating formatting quality with content quality) rather than generation-level reasoning errors. Plank (2022) argues that disagreement in human annotations carries meaningful signal about item difficulty; we demonstrate the converse—that *agreement* can itself be noise, a Shared Illusion driven by heuristics rather than genuine shared understanding of quality.

6 Conclusion

Through 105,600 evaluation instances, we formalize **Evaluation Illusion**—the phenomenon where LLM judges anchor scores on shared surface heuristics despite generating sophisticated critiques. Three structural mechanisms sustain this illusion: the **Resolution Paradox** ($\rho_{\text{model}} = 0.989 \gg \bar{r}_{\text{sample}} = 0.72 > \overline{\text{ICC}} = 0.67$), the **Rubric Commensurability Problem** (rubric structure accounts for 62% of agreement), and a negative quality–agreement correlation ($\rho = -0.513$) confirming the illusion is strongest where alignment signals matter most.

MERG's knowledge-grounded rubric generation recalibrates evaluation toward substance ($\Delta_K = -0.14$ to -0.22, $d = 0.97$–1.42), with domain-selective effects that rule out the noise hypothesis: agreement increases in codified domains but decreases in subjective ones. We advocate dynamically enriching evaluation rubrics with domain knowledge and explicitly distinguishing structural from substantive consensus to produce more meaningful alignment signals.

Limitations

Our study has several limitations. First, our evaluation is conducted on creative and professional writing tasks via WritingBench; the generalizability of Evaluation Illusion to exact-answer domains (e.g., code generation, mathematical reasoning) warrants further investigation. Second, all three evaluators are proprietary frontier models (Claude, Gemini, GPT); the findings may not transfer to open-source or fine-tuned judge models (Zhu et al., 2025). Third, we lack human ground-truth annotations and therefore cannot claim which evaluation method is absolutely more *accurate*; only that knowledge injection deconstructs the illusion of agreement. Fourth, the full MERG ablation was limited to two representative models (DeepSeek-R1 and Qwen3-235B) due to computational cost; a broader deployment across all 32 models would strengthen generalizability claims. Fifth, the RLAIF downstream experiment is preliminary, using a single policy model (Llama-3-8B-Instruct) on a subset of tasks; larger-scale alignment experiments are needed to confirm the observed mitigation of reward overoptimization. Finally, deeper evaluation is not necessarily more agreed-upon; the irreducible disagreement in subjective domains (Literature $\Delta_K = -0.06$) reflects genuine differences in aesthetic judgment, not evaluation failure.

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 (Claude, Gemini, GPT) 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 acknowledge 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 real evaluation scores from our experiments (Sample #501) applied to an LLM-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, and 1 others. 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. *arXiv preprint arXiv:2210.10760*.

Daniel Kahneman. 2011. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, New York.

Seungone Kim, Jamin Shin, Yejin Choi, 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, Kellie Lu, Thomas Mesnard, Colton Bishop, Victor Carbune, and Abhinav Rastogi. 2023. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*.

Yukyung Lee, Joonghoon Kim, Sagong Kim, Jooyoung Yi, Dongha Kim, and Jongwoo Lee. 2025. Checkeval: A reliable llm-as-a-judge framework for evaluating text generation using checklists. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, and 1 others. 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. *GitHub repository*.

Zongjie Li, Chaozheng Wang, Pingchuan Ma, Daoyuan Wu, Shuai Wang, Cuiyun Gao, and Yang Liu. 2024c. Split and merge: Aligning position biases in large language model 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 Halber, Luyu Deng, Romal Thoppilan, and Hannaneh Hajishirzi. 2023. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36.

Aditya Pathak and 1 others. 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. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*.

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.

Rafael Rafailov, Shohei Yamaguchi, Neel Chitpin, Philip Eddy, Emiliano Blagoja, Siyuan Zhuang, Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, and Stefano Ermon. 2024. Scaling laws for reward model overoptimization in direct alignment algorithms. *arXiv preprint arXiv:2406.02900*.

Delip Rao and Chris Callison-Burch. 2026. Autorubric: A unified framework for rubric-based LLM evaluation. *arXiv preprint arXiv:2603.00077*.

MohammadHossein Rezaei, Robert Vacareanu, Zihao Wang, and 1 others. 2025. Online rubrics elicitation from pairwise comparisons. *arXiv preprint arXiv:2510.07284*.

Liang Shan, Kaicheng Shen, Wen Wu, Zhenyu Ying, Chaochao Lu, and 1 others. 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, Arezou Nazemian, Haoran Xu, Dongfu Xiao, Banghua Zhu, Lianmin Zheng, Joseph E Gonzalez, 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, Feifan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023. Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926*.

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

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 Mun, Hwaran Lee, Sangdoo Yun, 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, 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, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 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 evaluator. 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: what to overlook?
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 pattern. All 8 Base models fall within the bottom tier (ranks 20 to 32), occupying 8 of the 13 lowest positions with high agreement ($\bar{r} = 0.81$), while the top-ranked Thinking models show the lowest agreement ($\bar{r} = 0.76$).

Table 9: **Complete 32-model rankings at t=0.0.** Quality and agreement correlation: $\rho = -0.513$, $p = 0.003$. Top-quality models (Thinking tier) consistently show the *lowest* evaluator agreement, confirming the Evaluation Illusion hypothesis.

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-Qwen32B	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-Llama8B	4.22	.845
25	Qwen3-30B-Base	3.91	.894
26	DeepSeek-R1-Distill-Qwen7B	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

The only exception is Llama-3.1-8B-Base (rank 32, $r = 0.595$), whose outputs are so incoherent that evaluators disagree on *how* to penalize them.

C Extended Case Studies

We present three cases from the maximum-disagreement analysis under MERG, illustrating distinct patterns of Evaluation Illusion. All scores are from MERG evaluation of DeepSeek-R1-0528 outputs.

Case 1: Patent Disclosure (Sample #45). *Spread: 6.80* (Gemini: 10.0, Claude: 6.4, GPT: 3.2)

Gemini praised "credible and specific details" and "exceptional performance across all dimensions." **GPT**, after knowledge activation, identified that the response misused Zero-Knowledge Proof (ZKP) cryptographic concepts and fabricated implementation details not grounded in the source paper. **Claude** detected structural merit but flagged technical inaccuracies. This case exemplifies *Concept Auditing vs. Keyword Matching*, where Gemini

matched technical keywords without verifying their correctness.

Case 2: Creative Fiction (Sample #739). *Spread: 6.70* (Gemini: 9.7, GPT: 7.1, Claude: 3.0)

Gemini awarded 9.7 for "exceptional thematic synthesis" and "a masterful integration of all prompt elements." **Claude** recognized that the model had generated a sophisticated structural outline rather than the requested narrative prose, noting: "An outline, however sophisticated, cannot score above 5 to 6 because it fails to deliver the requested deliverable." **GPT** acknowledged the conceptual strength but noted the writing remained at a planning stage. This case exemplifies *Substance vs. Outline*, where Gemini evaluated the *concept* while Claude evaluated the *deliverable*.

Case 3: Business Pitch Deck (Sample #501). *Spread: 6.20* (Gemini: 9.9, Claude: 6.5, GPT: 3.7)

Gemini scored 9.9, praising "a masterful understanding of the core task" and "a strategic masterpiece." **GPT**, after knowledge grounding, identified that the pitch entirely ignored China's "Double Reduction" policy (which banned for-profit K-12 tutoring), that financial projections were internally contradictory (ARPU calculations did not match claimed totals), and that it represented "piling numbers without building a model." **Claude** similarly detected regulatory blindness and placeholder data.

Pattern Summary. Across 100 samples, Gemini scored higher than both other evaluators in 90% of cases (mean 9.0 vs. Claude 7.1, GPT 6.9). The mean pairwise absolute difference was Claude/Gemini 2.32, Claude/GPT 1.10, Gemini/GPT 2.34. Of 100 samples, 78% showed a score spread >2 points, 41% >3 points, and 19% >4 points. This systematic pattern suggests that Gemini is most susceptible to Evaluation Illusion under MERG, consistently rewarding surface presentation even after knowledge activation.

D Sample-Level Agreement

Figure 5 presents the complete sample-level agreement landscape across all 32 evaluated models at 11 temperature settings ($t \in \{0.0, 0.1, \ldots, 1.0\}$). Four key observations emerge from this comprehensive view:

Temperature Invariance. Across all models, sample-level Pearson r remains remarkably stable as temperature varies. This rules out a simple explanation that disagreement arises from stochastic variation in evaluator outputs; even at t=0.0 (fully deterministic), agreement rarely exceeds $r = 0.80$. The flatness of the curves indicates that Evaluation Illusion is a *structural* property of how LLM evaluators process writing quality, not a consequence of sampling noise.

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 model tiers, suggesting that Claude and GPT share more similar internal evaluation heuristics, while Gemini employs a meaningfully different scoring strategy, consistent with the finding that Gemini is most susceptible to surface-level heuristic scoring (§3.1).

Model-Type Effects. Base models (light blue shading) tend to show *higher* sample-level agreement ($\bar{r} \approx 0.81$ at t=0.0) compared to Thinking models (dark shading, $\bar{r} \approx 0.76$). This inverse relationship between model quality and evaluator agreement supports the Resolution Paradox. Lower-quality outputs produce more obviously identifiable flaws that evaluators converge on, while high-quality outputs push evaluation into the ambiguous region where heuristic divergence dominates.

Outlier Models. Several models exhibit notably lower agreement: DeepSeek-R1-0528-Thinking and Qwen3-235B-A22B-Thinking (the two highest-quality Thinking models) show the lowest sample-level r values (≈ 0.55 to 0.65), reinforcing the quality and agreement inverse correlation documented in §3.2. Conversely, the weakest Base models (Llama-4-Scout-Base, Qwen2.5-7B-Base) show agreement approaching $r = 0.88$, where quality differences are unambiguous.

E Case Studies Under MERG

This section presents representative case studies illustrating how knowledge injection reshapes evaluator 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 near-perfect consensus to extreme divergence. Because many WritingBench tasks are in Chinese,

Figure 5: **Sample-level agreement is universally fragile.** Pearson r across 32 LLMs at 11 temperatures ($t \in [0.0, 1.0]$). Each subplot shows one model; lines represent evaluator pairs (**Claude/Gemini**, **Claude/GPT**, **Gemini/GPT**); background shading indicates model type (**Base**, **Instruct**, **Thinking**). The pattern is consistent: agreement is moderate ($r \approx 0.5$ to 0.8) and largely temperature-invariant, confirming that the Resolution Paradox (high model-level ρ, low sample-level r) is a universal property of LLM evaluation, not a model-specific artifact.

some evaluators produced Chinese-language reasoning; we present these verbatim and provide English translations (marked "translated from Chinese original").

E.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. All three evaluators independently identified the same strengths and weaknesses after knowledge activation.

- **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 evaluators also gave high scores (Brand Integration: 9.3, Persuasive Language: 9.3), but notably failed to penalize technical inaccuracies in material

claims (Technical Accuracy: only 4.7 in baseline). MERG's knowledge activation forced all three evaluators to independently discover and weight this deficiency, *yet they still converged* on a final score, demonstrating that agreement under knowledge grounding reflects genuine consensus rather than shared heuristics.

E.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 evaluators 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 evaluators 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 forced all three evaluators to verify claims against domain knowledge, independently reaching the same conclusion.

E.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 disclosure 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). The baseline consensus masked a fundamental question that MERG exposed: *should a patent disclosure be evaluated on its formal completeness or its legal viability?* Different evaluators, armed with the same domain knowledge, reached different but individually defensible conclusions, a case of irreducible evaluative pluralism.

E.4 Case D: The Formatting Trap

Task. Series A investment pitch deck for an EdTech startup 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 evaluators gave uniformly high scores (Strategic Positioning: 8.7, Content & Professionalism: 8.7, Financial Credibility: 7.3). The baseline consensus was built entirely on surface-level professionalism. MERG deconstructed this consensus by activating regulatory knowledge, revealing that evaluators weight domain violations very differently, from Gemini's format-first approach to GPT's substance-first stance.

E.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 evaluators 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，缺乏与之匹配的细节证据和具体分析：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 demonstrates MERG’s capacity to produce *calibrated* agreement even in subjective domains. Despite film criticism being inherently subjective (the Literature domain shows $\Delta_K = -0.06$ in Figure 4), knowledge activation anchored evaluators on verifiable criteria (factual accuracy and textual evidence) while preserving legitimate divergence on aesthetic judgments.

E.6 Summary of Patterns

Across these cases, three patterns emerge.

1. **Knowledge-grounded agreement is more meaningful**. When evaluators agree after MERG (Cases A, B, E), they cite the same specific evidence (material accuracy, data fabrication, factual errors) rather than generic praise for “professional tone” or “clear structure.”
2. **Disagreement reveals evaluation philosophy**. When evaluators disagree after MERG (Cases C, D), the disagreement is *interpretable*, reflecting different but defensible weightings of format vs. substance, structural completeness vs. legal validity.
3. **Baseline agreement is shallow**. In every case, baseline scores showed higher agreement than MERG scores, but this agreement was anchored on surface features that all evaluators could assess without domain expertise.

F Score Distributions

Figure 6 displays the score distributions for all 32 models across 3 evaluators at t=0.0, providing the most comprehensive view of the evaluation landscape.

Three-Tier Quality Hierarchy. The most striking feature is the robust separation of model tiers. 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. This ordering is recovered *unanimously* by all three evaluators (Cohen’s $d = 2.58$ for Base vs. Thinking), confirming that LLM evaluation remains valid for coarse-grained model comparison.

Score Compression in the High-Quality Region. Within the Thinking tier, score distributions become notably compressed, with the gap between the best and worst Thinking model being only $\sim$1.5 points (compared to $\sim$3.5 points within Base models). This compression is precisely the region where RLAIF reward models must discriminate between “good” and “great,” and where Evaluation Illusion is most damaging. The narrow dynamic range means that even small heuristic biases can invert model rankings at the sample level.

Evaluator-Specific Distribution Shapes. Across all models, evaluator-specific patterns

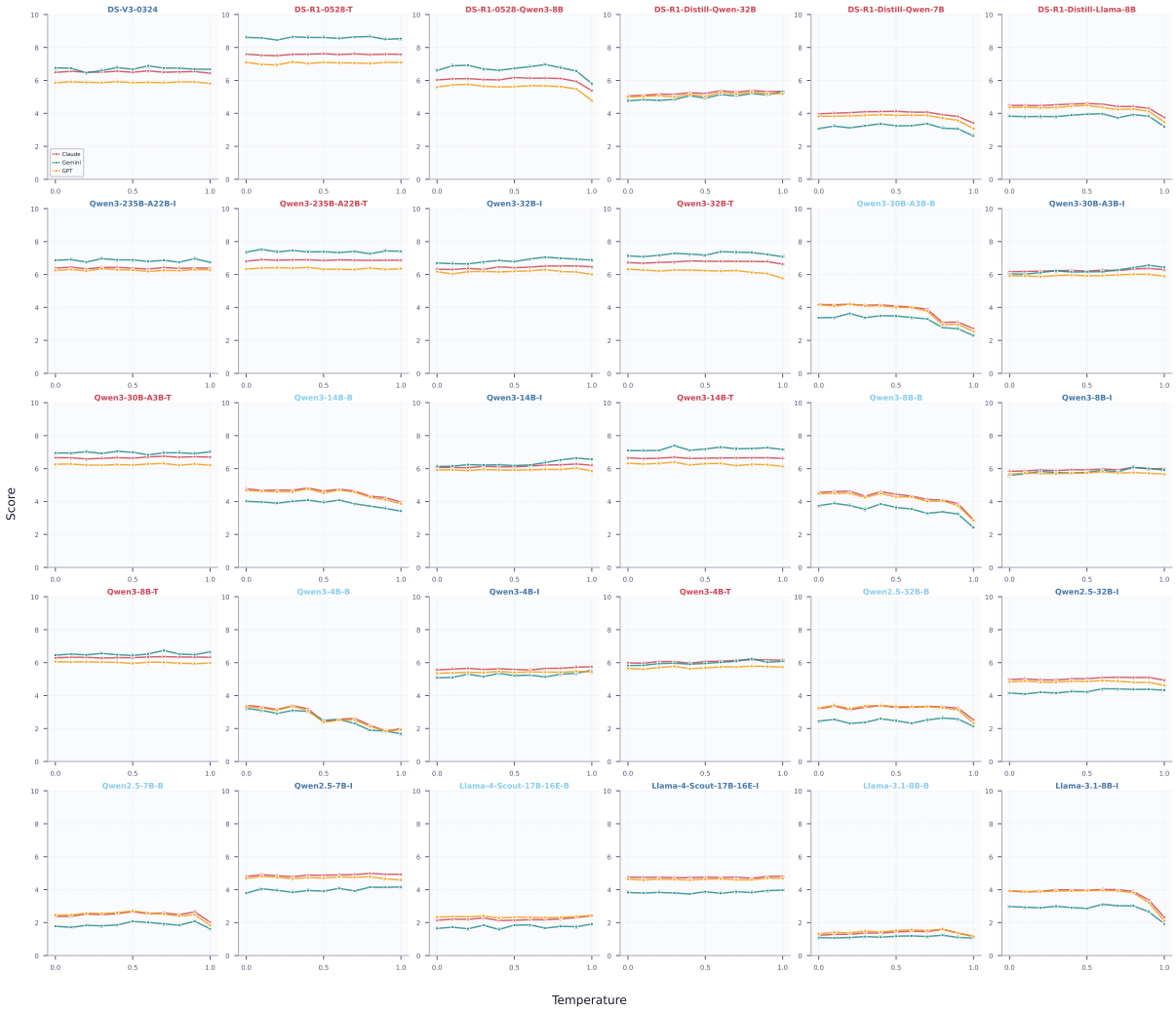

Figure 6: **System validity: the three-tier quality hierarchy is robust.** Score distributions for all 32 models across 3 evaluators at t=0.0. Despite the disagreement at the sample level documented throughout this paper, evaluators *unanimously* recover the Base $<$ Instruct $<$ Thinking quality ordering (Cohen's $d = 2.58$). This confirms that Evaluation Illusion does not invalidate LLM evaluation for coarse-grained model comparison; it specifically affects the fine-grained, per-sample signals that RLAIF reward models require.

are visible.

- **Gemini** (green) consistently produces the most right-skewed distributions, with a tendency toward high scores and less variance. This confirms the leniency bias documented in our main experiments. Gemini's distributions are shifted ∼0.5 to 1.0 points higher than Claude and GPT across all model tiers.
- **GPT** (blue) shows the widest variance, with heavier tails in both directions. This suggests GPT applies more differentiated scoring, producing stronger penalties for poor outputs and higher rewards for excellent ones.
- **Claude** (red) occupies a middle ground, with distributions that most closely approximate a normal shape. Claude's scores are the most calibrated in terms of centering, though it shares GPT's tendency to penalize specific failures more aggressively than Gemini.

Model-Size Effects. Within each tier, larger models generally receive higher mean scores, but the effect is modest ($\Delta\bar{s} \approx 0.3$ to 0.5 per order of magnitude in parameter count). The most notable exception is among Thinking models, where DeepSeek-R1-0528-Thinking and Qwen3-235B-A22B-Thinking achieve the highest absolute scores despite also exhibiting the lowest sample-level agreement (Figure 5), providing further evidence that the evaluation system is simultaneously *valid at the model level* and *unreliable at the sample level*.

Implications for RLAIF. The combination of (1) compressed score ranges in the high-quality

tier, (2) evaluator-specific scoring biases, and (3) low sample-level agreement creates a perfect storm for reward overoptimization. An RLAIF system training on these scores would receive reward signals that are dominated by evaluator idiosyncrasies rather than genuine quality differences, which is precisely the mechanism discussed in §4 (Implications for Reward Modeling).