# ScienceCritic: Empowering AI Review Systems with Non-objective Scientific Value Judgment

**Yaowen Hu[1,2,*], Qingbin Zeng[1], Ke Liang[2], Yafei Huang[1], Jianye Ju[1], Fengli Xu[1], Xinwang Liu[2], Yong Li[1]**
[1]Department of Electronic Engineering, Tsinghua University, Beijing, China
[2]College of Computer Science and Technology, National University of Defense Technology, Changsha, China
**Correspondence:** huyw26@mails.tsinghua.edu.cn

## Abstract

While LLMs offer a potential solution to the scalability crisis in academic peer review, they often prioritize linguistic fluency and objective metrics over the non-objective scientific value, which is crucial for scientists to identify high-impact or even award papers. To evaluate scientific value, we introduce ScienceCritic, which reformulates peer review as a structured causal chain that decomposes manuscripts across 4 dimensions derived from funding agency criteria (Strategic Scope, Gap Resolution, Intellectual Merit, and Broader Impacts according to NSF and NIH), and uses these scores to dictate subsequent tier decisions. To support ScienceCritic, we construct SCICRITIC-8K, a dataset of 8,696 conference submissions, each pairing the manuscript with its scores across the 4 dimensions, the corresponding rationales, and the final tier. To combat the sparsity of quantitative signals during training, our methodology introduces Ordinal-Constrained Fine-Tuning including (1) a Progressive Alignment Curriculum that first establishes the mapping from manuscript to scores and subsequently from scores to tier decisions, and (2) Triple-Objective optimization that enforces structure-aware token localization to anchor generative reasoning in numerical assessments. ScienceCritic achieves 72.9% overall accuracy and a 0.69 Spearman correlation with expert consensus, while better identifying high-impact research with 79.5% Oral and 40.0% Award recall, whereas the best AI Reviewer Agent only achieves 38.0% and 0.0% respectively.

## 1 Introduction

While empirical results and technical correctness offer objective metrics for evaluating research, assessing genuine scientific value remains an inherently non-objective endeavor. Yet, it is this very non-objective standard that distinguishes groundbreaking innovations from incremental improvements to existing methodologies (Liang et al., 2024). However, maintaining this standard via traditional peer review faces two main challenges: on one hand, the surge in submissions has created an unsustainable burden on reviewers; on the other, inherent cognitive biases often lead human reviewers to reject counter-intuitive, "non-consensus" ideas. While existing AI review systems aim to reduce this burden, they typically fixate on objective dimensions such as novelty and technical soundness and neglect the non-objective value judgments necessary for identifying high-impact research.

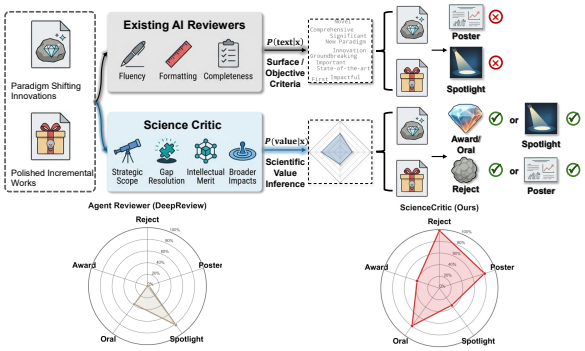


Figure 1: **Top:** Existing AI review systems ($P(\text{text}|x)$) optimize for surface-level objective criteria, failing to capture non-objective scientific value. **Bottom:** SCIENCECRITIC reformulates review as structured value inference ($P(\text{value}|x)$), decomposing scientific merit into 4 dimensions. This enables accurate identification of high-impact research: SCIENCECRITIC achieves **40%** Award and **79.5%** Oral recall vs. **0%** and **38%** for the best AI reviewer baseline.

Consequently, recent work has explored LLM-based assistants (Liang et al., 2024). Existing methodologies fall into two primary paradigms: (1) **Prompt-Centric Approaches**: Early works primarily utilized prompt engineering or in-context learning to guide general-purpose models in mimicking the structure and tone of professional human reviews (Zheng et al., 2023b; Li et al., 2025). (2) **System-Centric Approaches**: More recent endeav-

ors employ Multi-Agent Frameworks or Retrieval-Augmented Generation to simulate peer discussion or incorporate external knowledge, aiming to enhance the comprehensiveness and information density of feedback (Jin et al., 2024; Zhu et al., 2025).

Despite this, we identify three main limitations of existing AI review systems. **First**, constrained by the statistical priors of their training corpora, these systems favor incremental improvements that fit the probability distribution. Conversely, they often undervalue counter-intuitive paradigm breakthroughs as statistical outliers (Latona et al., 2024). **Second**, when deployed as multi-agent frameworks, AI reviewers typically converge to a groupthink consensus. This results in homogenization of critique and an inability to reliably identify high-value research. **Finally**, these systems exhibit severe positivity bias because they often avoid critical judgments and fail to recommend rejection even when papers contain serious flaws (Garg et al., 2025).

As illustrated in Figure 1, we introduce SCIENCECRITIC to reformulate automated peer review from unstructured text generation ($P(\text{text}|x)$) into a structured causal chain ($P(\text{value}|x)$). It departs from paradigms that optimize for linguistic plausibility. Instead, we decompose this value into 4 structural dimensions based on funding agency criteria (Strategic Scope, Gap Resolution, Intellectual Merit, Broader Impacts according to NSF and NIH). By structurally factorizing the inference process, we decouple the cognitive steps of peer review: the model is forced to anchor its qualitative reasoning to calibrated score intervals before deducing a final acceptance tier.

**Contributions.**

- We formalize peer review as Structured Value Inference, decomposing non-objective scientific value into 4 structural dimensions adapted from NSF and NIH criteria. To support this framework, we introduce SCICRITIC-8K, a dataset of 8,696 top-tier conference submissions paired with ground-truth consensus scores and rationales synthesized via a score-conditioned reasoning trajectory generation pipeline.
- We propose a **Triple-Objective Optimization** framework regulated by a **Progressive Alignment Curriculum**. To combat judgment signal sparsity, we enforce structure-aware token localization with ordinal constraints. This mechanism ensures that the model's text generation is driven by precise numerical representations, directly preventing autoregressive hallucinations and superficial stylistic mimicry.
- Experiments demonstrate the efficacy of our framework. SCIENCE CRITIC achieves an overall tier-adjudication accuracy of 72.9% and Spearman correlations above 0.63 across all 4 dimensions. It attains 98.5% Reject recall and 40.0% Award recall, identifying both severe methodological flaws and paradigm-shifting innovations.

## 2 Related work

### 2.1 Automated Peer Review

Early scholarly document processing treated peer review as localized auxiliary tasks, focusing on surface-level refinement such as grammatical correction and citation recommendation (Yuan et al., 2022; Rothe et al., 2021; Färber and Jatowt, 2020). Prompted by recent general-purpose LLMs (Liang et al., 2024), the paradigm shifted toward end-to-end generative reviewing. To elicit reliable academic assessments, recent frameworks fall into three main directions: imposing structured reasoning pipelines via dynamic prompting or hierarchical decision trees (Gao et al., 2024; Chang et al., 2025; Lu et al., 2024); injecting formal academic expertise through domain-specific instruction tuning and knowledge graph grounding (Zhu et al., 2025; Yu et al., 2024; Wang et al., 2020); and employing multi-agent deliberation protocols to simulate program committee dynamics and cross-examination (Wu et al., 2024; D'Arcy et al., 2024; Jin et al., 2024; Lu et al., 2025). Despite these enhancements, generative systems often focus on mimicking human style rather than making independent value judgments. Optimizing for stylistic alignment rather than systematic merit exacerbates sycophancy (Wei et al., 2023; Chen et al., 2024), causing models to validate safe, incremental work while failing to appreciate disruptive innovations.

### 2.2 Alignment and Ordinal Evaluation

Modern alignment of LLMs with human intent relies on RLHF and DPO variants (Christiano et al., 2017; Stiennon et al., 2020; Ouyang et al., 2022; Bai et al., 2022; Rafailov et al., 2023; Ethayarajh et al., 2024) to optimize for relative conversational

preferences (Wang et al., 2022). However, scientific peer review demands absolute evaluation against rigid rubrics (Casper et al., 2023). Applying pairwise preference optimization in specialized domains often leads to reward misspecification and superficial stylistic mimicry rather than learning a continuous quality representation (Gao et al., 2023; Perez et al., 2023; Sharma et al., 2023; Zhou et al., 2023). Furthermore, when adapted for quantitative scoring, standard Supervised Fine-Tuning (SFT) treats outputs as discrete categorical tokens (Zheng et al., 2023a), ignoring the inherent distance-aware ordinality of scientific judgment (Zhao et al., 2021). Although classical machine learning frequently uses ordinal regression (McCullagh, 1980; Gutiérrez et al., 2015; Niu et al., 2016), extending this concept to generative LLM vocabularies is less common. This difference is further hindered by signal sparsity: the handful of critical judgment tokens are statistically overwhelmed by hundreds of high-entropy rationale tokens during standard maximum likelihood estimation (Wang et al., 2023; Liu et al., 2024). Existing approaches either optimize for stylistic mimicry without grounding judgments in quantifiable metrics, or treat scores as flat categorical tokens that ignore ordinal structure. SCIENCE CRITIC addresses these gaps by reformulating review as Structured Value Inference over a 4-dimensional scientific merit space and introducing ordinal-constrained objectives that enforce calibrated, distance-aware scoring.

## 3 Methodology

### 3.1 Structured Value Inference

Standard instruction-tuning typically formulates automated peer review as a conditional density estimation problem. For a given manuscript $x \in \mathcal{X}$ and ground-truth review text $y \in \mathcal{Y}$, it minimizes the Negative Log-Likelihood (NLL) as shown in Eq. (1):

$$\mathcal{L}_{\text{SFT}} = -\sum_{t=1}^{|y|} \log p_\theta(y_t \mid x, y_{<t}), \qquad (1)$$

where $|y|$ denotes the sequence length, and $p_\theta(y_t \mid x, y_{<t})$ represents the autoregressive token-generation probability.

As illustrated in Figure 2(B), we reformulate the task as a Scientific Taste Alignment problem, transitioning from unstructured text generation to Structured Value Inference. The target output comprises a structured tuple $\tau = (\mathbf{s}, \mathbf{c}, z)$, where $\mathbf{s} \in \mathbb{R}^4$ represents a multi-dimensional value assessment, $\mathbf{c} \in \mathcal{C}$ is a reasoning rationale, and $z$ is a decision in a strictly ordered discrete space $\mathcal{Z}_{\text{ordered}}$. To reflect the structured causal chain of scientific judgment, we factorize the joint distribution $P_\theta(\tau \mid x)$ via the probability chain rule:

$$P_\theta(\tau \mid x) = \underbrace{P_\theta(z \mid \mathbf{s}, \mathbf{c})}_{\text{Decision}} \cdot \underbrace{P_\theta(\mathbf{c} \mid x, \mathbf{s})}_{\text{Reasoning}} \cdot \underbrace{P_\theta(\mathbf{s} \mid x)}_{\text{Quantification}}. \qquad (2)$$

Here, the reasoning rationale $\mathbf{c}$ is strictly anchored by the preceding quantitative assessment $\mathbf{s}$, and the final adjudication $z$ is conditionally independent of $x$ given $\mathbf{s}$ and $\mathbf{c}$.

To ensure $\mathbf{s}$ captures the non-objective scientific value distribution, we define its basis vectors using 4 structured evaluation dimensions derived from the NSF and NIH review criteria. We adopt these established criteria as they provide a domain-agnostic evaluation basis that isolates intrinsic scientific value from presentation quality. Detailed scoring rubrics and prompt definitions for dimension instantiation appear in Appendix C. The 4 dimensions are:

**Strategic Scope ($s_1$):** Evaluates Urgency and Relevance by assessing alignment with high-priority scientific challenges, distinguishing critical inquiries from marginal puzzles.

**Gap Resolution ($s_2$):** Measures Problem Closure by quantifying success in filling critical knowledge gaps.

**Intellectual Merit ($s_3$):** Assesses Fundamental Advancement and the potential to advance theoretical understanding within and across fields.

**Broader Impacts ($s_4$):** Proxies Societal Utility and the potential to benefit society beyond the immediate academic community.

### 3.2 Consensus Distillation

This work gathers an 8,696-paper dataset from OpenReview, spanning ICML, NeurIPS, and ICLR proceedings (2021–2025). For each submission, the pipeline extracts the manuscript text alongside official review metadata. Teacher Context $\mathcal{X}_{\text{oracle}}$ contains the full peer-review history to reflect human expert consensus. Student Context $\mathcal{X}_{\text{student}}$ restricts input to the abstract and introduction, which forces scientific value evaluation.

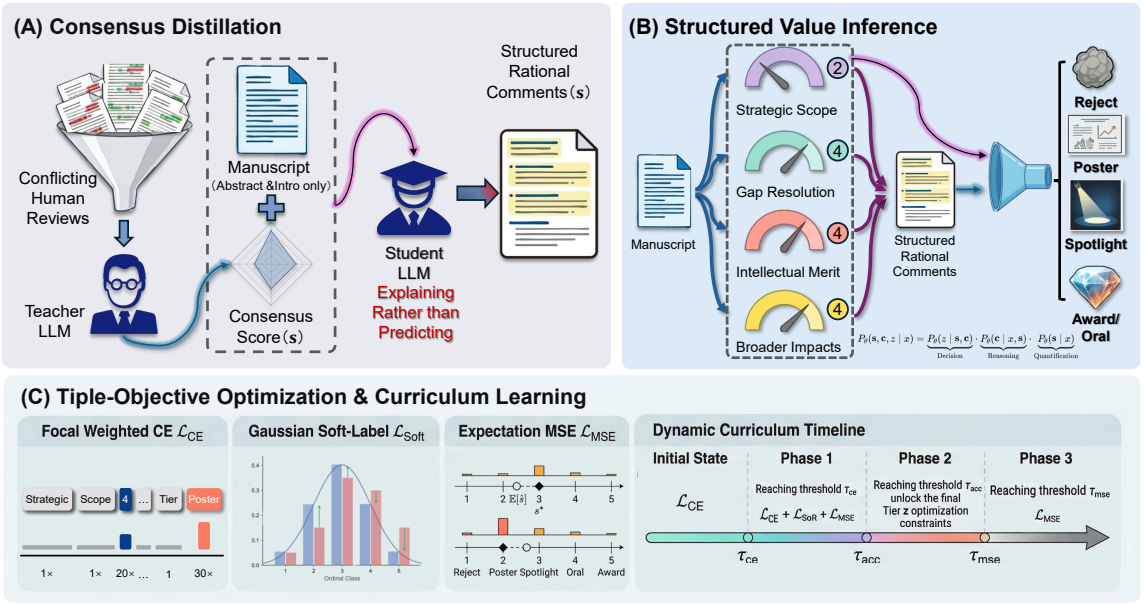


Figure 2: **Overview of the SCIENCE CRITIC framework. (A) Consensus Distillation:** A Teacher LLM distills conflicting human reviews into consensus scores ($\mathbf{s}^*$); a Student LLM then synthesizes explanatory rationales ($\mathbf{c}^*$) conditioned on $\mathbf{s}^*$ and the manuscript. **(B) Structured Value Inference:** Evaluation is factorized into a causal chain $P(\mathbf{s}, \mathbf{c}, z \mid x) = P(z \mid \mathbf{s}, \mathbf{c}) \cdot P(\mathbf{c} \mid x, \mathbf{s}) \cdot P(\mathbf{s} \mid x)$, quantifying the inherent non-objective scientific value across 4 structural dimensions before deducing the final acceptance tier. **(C) Triple-Objective Optimization & Curriculum Learning:** Focal $\mathcal{L}_{CE}$ isolates judgment tokens; $\mathcal{L}_{Soft}$ enforces ordinal consistency; $\mathcal{L}_{MSE}$ anchors numerical precision, all progressively unlocked via a curriculum timeline.

#### 3.2.1 Consensus Projection

Raw human reviews $\mathcal{R}$ often contain subjective noise and conflicting viewpoints. To extract the underlying scientific consensus, we employ DeepSeek-V3 as a projector to map this unstructured discourse onto our 4-dimensional value basis. Formally, we define a projection function $\Pi : \mathcal{R} \to [1, 5]^4$. The model acts as an aggregator, filtering out stylistic biases to estimate the consensus vector $\mathbf{s}^*$ following Eq. (3):

$$\mathbf{s}^* = \underset{\mathbf{s} \in [1,5]^4}{\operatorname{argmax}} P_{\text{Teacher}}(\mathbf{s} \mid \mathcal{R}; \Phi_{\text{schema}}). \quad (3)$$

Here, $\mathbf{s}^*$ is the consensus obtained by maximizing the teacher's likelihood over discourse $\mathcal{R}$, serving as the silver label for scientific value.

#### 3.2.2 Reasoning Trajectory Generation

To construct high-quality supervision, reasoning chains $\mathbf{c}$ must be causally derivable from the restricted student input $x_{\text{student}}$ while aligning with expert consensus scores $\mathbf{s}^*$. This work proposes score-conditioned reasoning trajectory generation to synthesize these rationales. The pipeline prompts the teacher model to function as a rationalizer, utilizing manuscript text $x_{\text{student}}$ and target scores $\mathbf{s}^*$ to produce a justification $\mathbf{c}^*$ in Eq. (4):

$$\mathbf{c}^* \sim P_{\text{Teacher}}(\mathbf{c} \mid x_{\text{student}}, \mathbf{s}^*). \quad (4)$$

The final dataset $\mathcal{D}_{\text{SFT}} = \{(x_{\text{student}}, \mathbf{c}^*, \mathbf{s}^*, z^*)\}$ comprises 4-tuples grounded in available context and aligned with scientific consensus.

### 3.3 Ordinal-Constrained Fine-Tuning

Using the distilled consensus dataset $\mathcal{D}_{\text{SFT}}$, we finetune the Science Critic to internalize expert adjudication criteria. Standard maximum likelihood estimation is ill-suited for this task, as the high-entropy rationale tokens ($\mathbf{c}^*$) statistically dominate the sparse judgment signals ($\mathbf{s}^*, z^*$), which constitute less than 1% of the sequence. We address this imbalance by introducing a Triple-Objective Optimization framework paired with a Progressive Alignment Curriculum, reweighting the loss landscape to enforce that the generative reasoning adheres to the quantitative assessment.

#### 3.3.1 Structure-Aware Token Localization

To apply judgment-specific losses to the structured components of $\tau$, we must locate the tokens corresponding to scores $\mathbf{s}^*$ and decision $z^*$ within the

**Algorithm 1** Progressive Alignment Curriculum

**Initialize:** Model $\theta$, Flags $E_{\text{score}} \leftarrow$ False, $E_{\text{tier}} \leftarrow$ False, $E_{\text{soft}} \leftarrow$ True
**Hyperparameters:** $\tau_{\text{ce}}$, $\tau_{\text{acc}}$, $\tau_{\text{mse}}$
**while** training not converged **do**
  *// 1. Forward Pass & Loss Composition*
  Cross-entropy loss: $\mathcal{L}_{\text{total}} \leftarrow \mathcal{L}_{\text{CE}}$
  Ordinal objective: $\mathcal{L}_{\text{Ord}}(\mathcal{I}) = \lambda_{\text{mse}}\mathcal{L}_{\text{MSE}}(\mathcal{I}) + \mathbb{I}_{[E_{\text{soft}}]} \cdot \mathcal{L}_{\text{Soft}}(\mathcal{I})$
  Active nodes: $\mathcal{I}_{\text{act.}} \leftarrow (\mathcal{I}_{\text{score}} \text{ if } E_{\text{score}}) \cup (\mathcal{I}_{\text{tier}} \text{ if } E_{\text{tier}})$
  **if** $\mathcal{I}_{\text{act.}} \neq \emptyset$ **then**
    $\mathcal{L}_{\text{total}} \leftarrow \mathcal{L}_{\text{total}} + \text{Avg}(\mathcal{L}_{\text{Ord}}(\mathcal{I}_{\text{act.}}))$
  **end if**
  Update $\theta$ via backpropagation $\nabla_\theta \mathcal{L}_{\text{total}}$
  *// 2. Curriculum State Updates*
  **if** $\text{Avg}(\mathcal{L}_{\text{CE}}) \leq \tau_{\text{ce}}$ **then**
    $E_{\text{score}} \leftarrow$ True
  **end if**
  **if** $\text{Acc}_{\mathbf{s}} \geq \tau_{\text{acc}}$ **then**
    $E_{\text{tier}} \leftarrow$ True
  **end if**
  **if** $\text{Avg}(\mathcal{L}_{\text{MSE}}) \leq \tau_{\text{mse}}$ **then**
    $E_{\text{soft}} \leftarrow$ False
  **end if**
**end while**

generated sequence. Let $\mathcal{X}$ denote the raw text sequence and $\mathcal{T} = \{t_1, \ldots, t_L\}$ be its tokenized representation. Pattern extraction via a regex operator $\mathcal{E}_{\text{regex}}$ identifies ground-truth character spans for all judgment values as defined in Eq. (5):

$$\mathcal{S}_{\text{gt}} = \{\sigma \subset \mathbb{N} \mid \sigma = [\text{start}, \text{end}), \mathcal{X}[\sigma] \in \{\mathbf{s}^*, z^*\}\}. \tag{5}$$

Here, regex operator $\mathcal{E}_{\text{regex}}$ extracts character-level intervals $\sigma \subset \mathbb{N}$ defined as half-open sets containing consensus scores and decisions. String slicing $\mathcal{X}[\sigma]$ isolates target values within the raw sequence.

To map these semantic spans to token positions, we must reconcile the character-level coordinates with their corresponding token positions. Accordingly, let $\mathcal{M} : \{1, \ldots, L\} \to \mathbb{N} \times \mathbb{N}$ be the tokenizer offset mapping, where $\mathcal{M}(i)$ returns the character interval token $t_i$ spans. Projecting semantic spans $\mathcal{S}_{\text{gt}}$ onto the token grid constructs the target index set $\mathcal{I}_{\text{val}}$ following Eq. (6):

$$\mathcal{I}_{\text{val}} = \{i \in [1, L] \mid \exists \sigma \in \mathcal{S}_{\text{gt}}, \mathcal{M}(i) \cap \sigma \neq \emptyset\}. \tag{6}$$

Specifically, $\mathcal{M}(i)$ projects token $t_i$ onto its character-space interval; $\mathcal{I}_{\text{val}}$ comprises indices with non-empty intersections (i.e., $\mathcal{M}(i) \cap \sigma \neq \emptyset$) with $\mathcal{S}_{\text{gt}}$, restricting gradient backpropagation to judgment tokens.

#### 3.3.2 Triple-Objective Optimization

Anchoring judgment tokens $\mathcal{I}_{\text{val}}$ allows us to combine training objectives that enforce ordinal consistency and quantitative precision. Three loss functions align the model's internal representations with community standards.

**Focal Weighted Cross-Entropy ($\mathcal{L}_{\text{CE}}$):** Standard SFT often suffers from the statistical dominance of rationale tokens, which can drown out the sparse signals of evaluative judgment. To counteract this, we implement a reweighting scheme to amplify gradients from judgment tokens, formulated as:

$$\mathcal{L}_{\text{CE}} = -\sum_t w_t \log p_\theta(x_t \mid x_{<t}). \tag{7}$$

**Ordinal Gaussian Soft-Label ($\mathcal{L}_{\text{Soft}}$):** Scientific judgment exhibits continuous ordinality, where the severity of an error depends on its distance from the ground truth. To capture this, we apply Gaussian smoothing $q(k) \propto \exp(-(k-y)^2/2\sigma^2)$ to the one-hot targets at $\mathcal{I}_{\text{val}}$, and minimize the divergence following Eq. (8):

$$\mathcal{L}_{\text{Soft}} = \sum_{i \in \mathcal{I}_{\text{val}}} D_{\text{KL}}(q_i \parallel p_\theta(\cdot \mid x_i)). \tag{8}$$

Within this objective, $D_{\text{KL}}$ quantifies the divergence between the model's predictive distribution and the smoothed target $q_i$.

**Expectation Regression ($\mathcal{L}_{\text{MSE}}$):** To further stabilize numerical outputs, we introduce an expected merit regression constraint. Defining the expected score as $\hat{s} = \mathbb{E}_{p_\theta}[k]$, the objective minimizes the squared discrepancy as shown in Eq. (9):

$$\mathcal{L}_{\text{MSE}} = (\hat{s} - s^*)^2. \tag{9}$$

#### 3.3.3 Progressive Alignment Curriculum

We formalize this causal dependency through a Progressive Alignment Curriculum, as detailed in Algorithm 1, which manages training dynamics across three distinct phases:

Initially, a structural warm-up stabilizes the model's language generation by optimizing on the base cross-entropy objective ($\mathcal{L}_{\text{CE}}$). Once the average $\mathcal{L}_{\text{CE}}$ falls below a stability threshold ($\tau_{\text{ce}}$), the

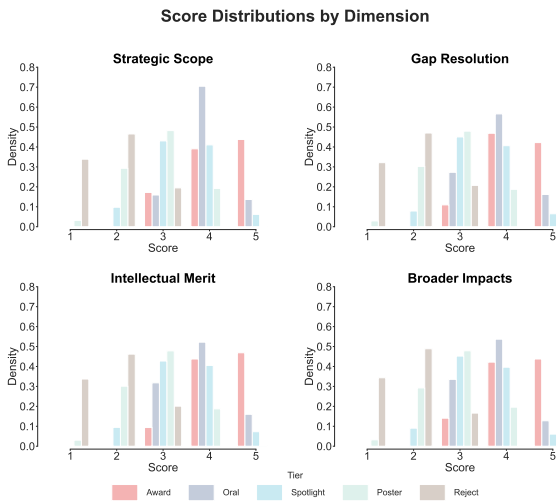


Figure 3: Four-dimensional consensus score distributions, showing monotonically increasing values from Reject to Award across all criteria.

curriculum activates additional value constraints ($\mathcal{L}_{\text{Soft}} + \mathcal{L}_{\text{MSE}}$) for the intermediate score tokens $\mathcal{I}_{\text{score}}$, ensuring that linguistic convergence precedes numerical calibration. Furthermore, to enforce the cognitive hierarchy of scientific judgment, the optimization of the final decision token $\mathcal{I}_{\text{tier}}$ is temporarily restricted to the base $\mathcal{L}_{\text{CE}}$ objective. The numerical constraints for $\mathcal{I}_{\text{tier}}$ are unlocked only after the model demonstrates sufficient value assessment competence, measured by the empirical score accuracy exceeding a predefined threshold ($\tau_{\text{acc}}$). Finally, upon reaching numerical convergence, when the mean squared error drops below $\tau_{\text{mse}}$, a precision refinement phase globally disables $\mathcal{L}_{\text{Soft}}$.

# 4 Experiments

## 4.1 Experimental Setup

We evaluate on SCICRITIC-8K using a stratified test set of 820 manuscripts (200 each for Reject, Poster, Spotlight, Oral; 20 Award), with the remaining 7,876 allocated for training. We base SCIENCE CRITIC on Qwen2.5-7B-Instruct (Qwen Team, 2024) via LoRA-based parameter-efficient SFT. General-purpose baselines include OpenAI o1 (OpenAI, 2024), GPT-4o (OpenAI, 2023), Claude 3.5 Sonnet (Anthropic, 2024), DeepSeek-V3 (DeepSeek-AI, 2024), and Qwen2.5 at 7B, 32B, and 72B scales (Qwen Team, 2024). Specialized academic frameworks include Reviewer2 (Gao et al., 2024), AgentReview (Jin et al., 2024), DeepReview (Zhu et al., 2025), and TreeReview (Chang et al., 2025). All models are tested under Zero-shot, Few-shot, Agentic, and Agentic+Few-shot configurations (detailed in Appendix G). Value estimation is measured by Spearman correlation ($r_s$); final adjudication by overall Accuracy ($Acc$) and Per-Tier Recall ($Rec_k$). Full baseline configurations, training hyperparameters, and metric definitions are in Appendix B.

## 4.2 Consensus Score Validation

To validate the reliability of silver labels, we verify that the teacher-generated scores monotonically stratify manuscripts across decision tiers. As shown in Figure 3, score distributions align with tier prestige. For example, the mean *Intellectual Merit* progressively rises from 1.86 ($\pm$0.72) for "Reject" to 4.38 ($\pm$0.66) for "Award", a trend universally observed across all dimensions (see Table 4 in Appendix E for full statistics).

Non-parametric Kruskal-Wallis tests (Kruskal and Wallis, 1952) confirm global statistical significance across all 4 criteria ($H > 2110, p < 0.0001$), and post-hoc Mann-Whitney U tests (Mann and Whitney, 1947) confirm that every adjacent tier transition is statistically significant (Figure 6 and Figure 7 in Appendix E). We conduct a human validation study with 15 volunteers, yielding an overall Spearman $\rho = 0.812$ between human annotations and the synthesized consensus labels, with per-dimension $\rho$ ranging from 0.758 to 0.844 (see Appendix J for details).

## 4.3 Comparison Study

We first evaluate whether off-the-shelf LLMs can perform reliable peer review. As shown in Figure 4, while ground truth scores span the full 1–5 Likert scale, baselines consistently cluster predictions at median values and virtually never assign low scores ($\leq 2.0$), making their tier-wise mean trajectories nearly flat and leaving them unable to distinguish a rejected paper from an award winner.

We also evaluate if advanced prompting reduces this bias. Appendix G.2 shows that few-shot examples and agentic reasoning shift predicted distributions only marginally across all LLM families; the central tendency bias persists, and tail tiers (Reject, Award) remain under-recalled. As shown in Table 1, all general-purpose baselines yield Spearman $r_s < 0.3$ across all 4 dimensions, and even specialized academic frameworks such as Reviewer2 and DeepReview fail to outperform zero-shot inference.

SCIENCE CRITIC addresses both issues. It

Table 1: Evaluation on SCICRITIC-8K (820 manuscripts). Best in **bold**; second-best underlined.

| Model / Framework | Mode | Value Estimation ($r_s \uparrow$) | | | | Overall | Per-Tier Recall (% ↑) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Scope | Gap | Merit | Impact | Acc | Rej. | Post. | Spot. | Oral | Awd. |
| *General-Purpose Open-Weight Models* | | | | | | | | | | | |
| Qwen2.5-7B (Qwen Team, 2024) | Zero-shot | 0.064 | 0.072 | 0.025 | −0.008 | 0.250 | 0.0 | 0.0 | 63.0 | 39.5 | 0.0 |
| | Few-shot | 0.008 | 0.031 | 0.003 | −0.034 | 0.238 | 0.0 | 9.5 | 67.5 | 20.5 | 0.0 |
| | Agentic | −0.013 | 0.083 | 0.053 | −0.055 | 0.249 | 0.0 | 0.0 | 64.0 | 38.0 | 0.0 |
| | Agentic + Few-shot | 0.027 | 0.098 | −0.024 | −0.038 | 0.232 | 0.0 | 10.5 | 66.0 | 18.5 | 0.0 |
| Qwen2.5-72B (Qwen Team, 2024) | Zero-shot | 0.046 | −0.010 | −0.006 | −0.096 | 0.254 | 0.0 | 0.0 | 81.0 | 23.0 | 0.0 |
| | Few-shot | −0.027 | 0.094 | −0.021 | −0.034 | 0.252 | 0.0 | 1.0 | 72.5 | 30.0 | 0.0 |
| | Agentic | −0.015 | 0.030 | 0.007 | −0.002 | 0.245 | 0.0 | 0.0 | **99.5** | 1.0 | 0.0 |
| | Agentic + Few-shot | 0.051 | 0.031 | 0.032 | −0.063 | 0.244 | 0.0 | 0.5 | 79.0 | 20.5 | 0.0 |
| DeepSeek-V3 (DeepSeek-AI, 2024) | Zero-shot | 0.037 | 0.078 | 0.058 | −0.007 | 0.288 | 1.0 | 34.5 | 68.5 | 14.0 | 0.0 |
| | Few-shot | −0.004 | 0.108 | 0.062 | −0.069 | 0.268 | 1.5 | 42.0 | 53.0 | 13.5 | 0.0 |
| | Agentic | 0.062 | 0.029 | 0.072 | −0.062 | 0.272 | 0.0 | 18.5 | 82.5 | 9.5 | 10.0 |
| | Agentic + Few-shot | 0.036 | 0.089 | 0.042 | −0.096 | 0.266 | 2.0 | 17.5 | 72.0 | 17.0 | 5.0 |
| *General-Purpose Proprietary Models* | | | | | | | | | | | |
| Claude 3.5 Sonnet (Anthropic, 2024) | Zero-shot | 0.029 | −0.043 | 0.072 | −0.021 | 0.181 | 0.0 | 2.5 | 71.5 | 0.0 | 0.0 |
| | Few-shot | −0.016 | −0.007 | 0.019 | 0.013 | 0.229 | 0.0 | 15.5 | 65.0 | 13.5 | 0.0 |
| | Agentic | 0.010 | 0.084 | −0.080 | 0.066 | 0.232 | 0.0 | 11.5 | 82.5 | 1.0 | 0.0 |
| | Agentic + Few-shot | −0.004 | 0.043 | −0.050 | 0.016 | 0.248 | 1.0 | 15.5 | 78.5 | 6.5 | 0.0 |
| GPT-4o (OpenAI, 2023) | Zero-shot | 0.053 | 0.030 | 0.098 | −0.040 | 0.248 | 0.0 | 1.5 | 97.5 | 2.5 | 0.0 |
| | Few-shot | 0.080 | 0.042 | 0.206 | 0.042 | 0.255 | 1.0 | 10.0 | 84.0 | 9.5 | 0.0 |
| | Agentic | 0.016 | 0.095 | 0.172 | −0.001 | 0.256 | 0.0 | 7.5 | 95.5 | 1.0 | 10.0 |
| | Agentic + Few-shot | 0.082 | 0.141 | 0.208 | −0.023 | 0.244 | 0.0 | 5.5 | 79.0 | 15.5 | 0.0 |
| OpenAI o1 (OpenAI, 2024) | Zero-shot | 0.021 | 0.062 | 0.185 | −0.002 | 0.279 | 0.5 | 22.5 | 86.0 | 5.5 | 0.0 |
| | Few-shot | 0.057 | 0.099 | 0.166 | 0.008 | 0.261 | 2.5 | 28.5 | 61.5 | 14.5 | 0.0 |
| | Agentic | −0.005 | 0.075 | 0.126 | −0.039 | 0.252 | 0.0 | 39.0 | 63.0 | 1.5 | 0.0 |
| | Agentic + Few-shot | 0.033 | 0.081 | 0.135 | −0.065 | 0.245 | 0.5 | 33.0 | 54.0 | 13.0 | 0.0 |
| *Specialized Academic Evaluation Frameworks* | | | | | | | | | | | |
| Reviewer2 (Gao et al., 2024) | Agentic | 0.006 | −0.044 | −0.036 | 0.013 | 0.261 | 0.0 | 16.0 | 83.5 | 7.5 | 0.0 |
| AgentReview (Jin et al., 2024) | Agentic | 0.017 | −0.002 | 0.030 | 0.034 | 0.238 | 0.0 | 7.0 | 71.5 | 19.0 | 0.0 |
| DeepReview (Zhu et al., 2025) | Agentic | 0.061 | 0.001 | 0.077 | 0.040 | 0.299 | 0.0 | 2.5 | 82.0 | 38.0 | 0.0 |
| TreeReview (Chang et al., 2025) | Agentic | −0.063 | −0.024 | −0.104 | 0.005 | 0.270 | 0.0 | 22.5 | 82.0 | 6.0 | 0.0 |
| **SCIENCE CRITIC (Ours)** | **Agentic** | **0.694** | **0.677** | **0.659** | **0.635** | **0.729** | **98.5** | **81.0** | 36.0 | **79.5** | **40.0** |

Table 2: Ablation results. **Part I:** All 8 combinations of Triple-Objective components (absent Focal $\mathcal{L}_{\text{CE}}$ = standard NLL). **Part II:** Ablating the inference-time Scoring Rubric (renders $r_s$ undefined).

| Objectives | | | Value Estimation ($r_s \uparrow$) | | | | Overall | Per-Tier Recall (% ↑) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Focal $\mathcal{L}_{\text{CE}}$ | $\mathcal{L}_{\text{Soft}}$ | $\mathcal{L}_{\text{MSE}}$ | Scope | Gap | Merit | Impact | Acc | Rej. | Post. | Spot. | Oral | Awd. |
| *Part I: Training Stage* | | | | | | | | | | | | |
| ✓ | ✓ | ✓ | **0.694** | **0.677** | **0.659** | **0.635** | **0.729** | **98.5** | **81.0** | **36.0** | **79.5** | **40.0** |
| ✓ | ✓ | | 0.620 | 0.601 | 0.604 | 0.548 | 0.634 | 74.0 | 72.5 | 35.5 | 76.0 | 20.0 |
| ✓ | | ✓ | 0.552 | 0.488 | 0.501 | 0.442 | 0.660 | 92.5 | 72.0 | 29.5 | 73.5 | 30.0 |
| | ✓ | ✓ | 0.530 | 0.498 | 0.477 | 0.467 | 0.624 | 88.0 | 70.0 | 35.5 | 60.0 | 25.0 |
| ✓ | | | 0.361 | 0.421 | 0.373 | 0.318 | 0.551 | 79.5 | 66.0 | 28.0 | 51.0 | 15.0 |
| | ✓ | | 0.385 | 0.344 | 0.348 | 0.328 | 0.461 | 69.0 | 58.0 | 21.5 | 39.5 | 10.0 |
| | | ✓ | 0.311 | 0.395 | 0.327 | 0.341 | 0.526 | 80.0 | 61.5 | 26.0 | 47.0 | 10.0 |
| | | | 0.264 | 0.259 | 0.293 | 0.192 | 0.374 | 55.0 | 49.5 | 19.5 | 29.5 | 0.0 |
| *Part II: Test Stage* | | | | | | | | | | | | |
| w/o Scoring Rubric | | | - | - | - | - | 0.493 | 91.0 | 59.5 | 21.5 | 30.0 | 0.0 |

achieves $r_s > 0.63$ on all 4 dimensions, outperforming the best baseline in value estimation. For tier adjudication, baselines concentrate predictions on middle tiers, resulting in near-zero Reject recall and degraded Award recall; in contrast, SCIENCE CRITIC maintains $Rec_{\text{Rej}} = 98.5\%$ and $Rec_{\text{Awd}} = 40.0\%$ (4× the best baseline), as shown in Figure 5. Detailed confusion matrices are pro-

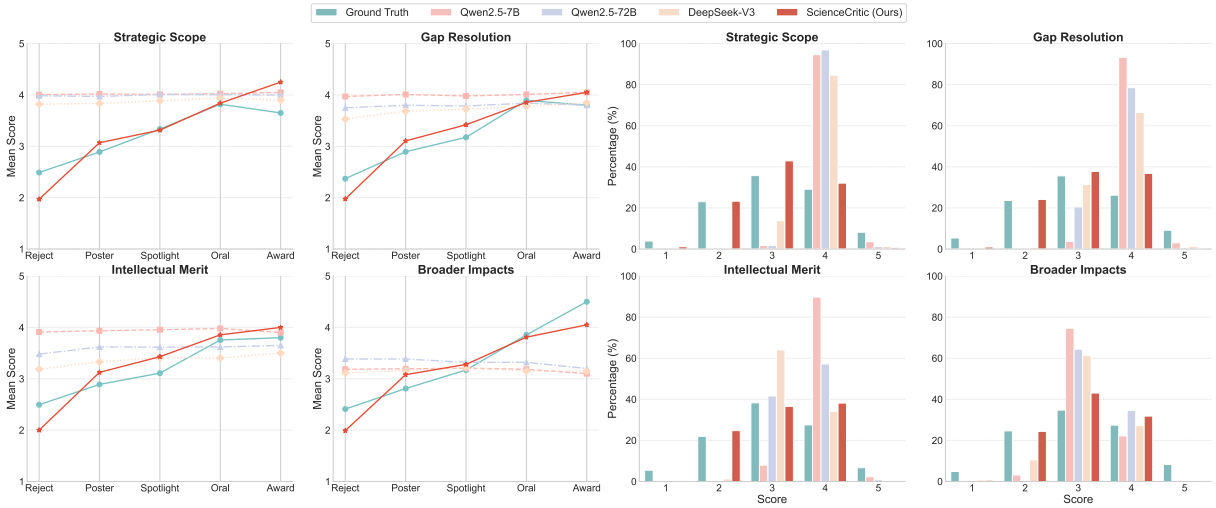


Figure 4: Distributional biases of baseline LLMs vs. SCIENCE CRITIC. **Left:** Tier-wise mean trajectories show that baselines produce nearly flat scores across tiers, while SCIENCE CRITIC closely tracks the ground truth's steep quality progression. **Right:** Score distributions reveal that baselines collapse around median values (3–4), whereas SCIENCE CRITIC restores the natural variance of the full scoring range.

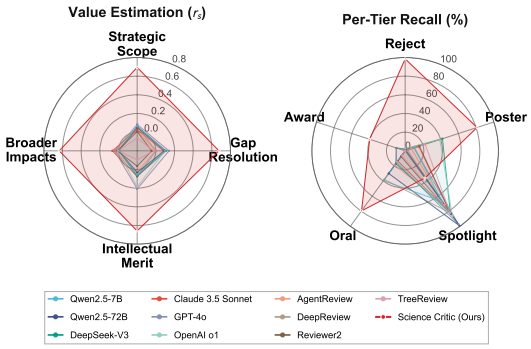


Figure 5: **Left:** Spearman $r_s$ across 4 dimensions. **Right:** Per-Tier Recall across five tiers. Baselines inflate middle-tier recall via central tendency bias while collapsing at the tails (Reject/Award).

vided in Appendix G.3.

### 4.4 Ablation Studies

We ablate the contributions of our proposed architecture, focusing on the Triple-Objective Optimization framework (Section 3.3.2) and the inference-time Scoring Rubric. Results are in Table 2.

**Part I** shows that all three objectives are necessary and play distinct roles. **Focal** $\mathcal{L}_{\text{CE}}$ is the most critical: removing it causes Strategic Scope $r_s$ to drop from 0.694 to 0.530 and Acc from 72.9% to 62.4%, as sparse judgment tokens are statistically overwhelmed by rationale tokens. $\mathcal{L}_{\text{Soft}}$ primarily stabilizes monotonic rank alignment: without it, Strategic Scope $r_s$ falls to 0.552 and $Rec_{\text{Awd}}$ drops to 30.0%. $\mathcal{L}_{\text{MSE}}$ anchors tail-tier calibration: its removal collapses $Rec_{\text{Awd}}$ to 20.0% and $Rec_{\text{Rej}}$ to 74.0%. Base SFT, which removes all three objectives, degrades to $r_s \approx 0.26$ and $Rec_{\text{Awd}} = 0\%$. **Part II** shows that removing the Scoring Rubric at test time drops Acc from 72.9% to 49.3% and $Rec_{\text{Awd}}$ to 0%, confirming that explicit dimensional axes are required to prevent reasoning hallucinations during adjudication.

## 5 Conclusion

In this work, we introduced SCIENCE CRITIC, a framework that reformulates automated peer review from unstructured text generation to Structured Value Inference. To address the sycophancy and preference for incremental improvements common in general-purpose LLMs, we constructed the SCICRITIC-8K dataset via a score-conditioned reasoning trajectory generation pipeline. By optimizing the model through a Progressive Alignment Curriculum and ordinal-constrained objectives, we enforce a process where quantitative judgments are derived from logically grounded reasoning. Our results indicate that advanced prompt engineering and multi-agent frameworks cannot fully correct the distributional biases in baseline models. On the contrary, SCIENCE CRITIC restores the natural variance of the evaluation space by anchoring judgments through our ordinal-constrained objectives and Progressive Alignment Curriculum.

## Limitations

The consensus distillation pipeline currently relies on a single teacher model (DeepSeek-V3). Different teacher models may yield varying consensus scores due to their distinct internal biases. Additionally, as our framework is specifically designed to assess non-objective scientific value, it intentionally focuses on the contribution claims articulated in the abstract and introduction. Consequently, verification of objective experimental correctness, reproducibility of results, and methodological soundness in later sections falls outside the current scope of this work.

## Ethics Statement

This research focuses on the automated evaluation of public scientific manuscripts and is designed to reflect systematic criteria aligned with human consensus. The SCICRITIC-8K dataset constructed for this study consists entirely of publicly available scientific manuscripts and official consensus metadata collected from open academic venues (ICLR, NeurIPS, ICML). The data contains no personally identifying information beyond public author identities or any offensive content.

This research did not involve human subjects, paid crowdworkers, or any behavioral experiments, and therefore does not require ethics review board approval. The computational experiments, including hyperparameter settings and statistics for data splits, are fully detailed to ensure reproducibility. Large Language Models were strictly utilized for language polishing, grammar correction, and proofreading of the manuscript drafted by the human authors.

## References


Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. Technical report, Anthropic. Anthropic Technical Report.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli Ganguli, Tom Henighan, and 1 others. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, and 1 others. 2023. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*.

Yuan Chang, Ziyue Li, Hengyuan Zhang, Yuanbo Kong, Yanru Wu, Hayden Kwok-Hay So, Zhijiang Guo, Liya Zhu, and Ngai Wong. 2025. Treereview: A dynamic tree of questions framework for deep and efficient llm-based scientific peer review. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15662–15693.

Daiwei Chen, Yi Chen, Aniket Rege, and Ramya Korlakai Vinayak. 2024. Modeling the plurality of human preferences via ideal points. In *ICML 2024 Workshop on Models of Human Feedback for AI Alignment*.

Xiaoshu Chen, Sihang Zhou, Ke Liang, and Xinwang Liu. 2025. Distilling reasoning ability from large language models with adaptive thinking. *IEEE Transactions on Neural Networks and Learning Systems*. ArXiv:2404.09170.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Mike D'Arcy, Tom Hope, Larry Birnbaum, and Doug Downey. 2024. MARG: Multi-agent review generation for scientific papers. *arXiv preprint arXiv:2401.04259*.

DeepSeek-AI. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. 2024. KTO: Model alignment as prospect theoretic optimization. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*.

Michael Färber and Adam Jatowt. 2020. Citation recommendation: Approaches and datasets. *International Journal on Digital Libraries*, 21(4):375–405.

Leo Gao, John Schulman, and Jacob Hilton. 2023. Scaling laws for reward model overoptimization. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*.

Zhaolin Gao, Kianté Brantley, and Thorsten Joachims. 2024. Reviewer2: Optimizing review generation through prompt generation. *Preprint*, arXiv:2402.10886.

Madhav Krishan Garg, Tejash Prasad, Tanmay Singhal, Chhavi Kirtani, Murari Mandal, and Dhruv Kumar. 2025. Revieweval: An evaluation framework for ai-generated reviews. *Preprint*, arXiv:2502.11736.

Pedro Antonio Gutiérrez, Maria Perez-Ortiz, Javier Sanchez-Monedero, Francisco Fernandez-Navarro, and Cesar Hervas-Martinez. 2015. Ordinal regression methods: Survey and experimental study. *IEEE*

*Transactions on Knowledge and Data Engineering*, 28(1):127–146.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Namgyu Ho, Laura Schmid, and Se-Young Yun. 2023. Large language models are reasoning teachers. *arXiv preprint arXiv:2212.10071*.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alexander Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8003–8017.

Yiqun Jin, Qinglin Zhao, Yibo Wang, Huaping Chen, Kael Zhu, Yang Xiao, and Jian Wang. 2024. Agentreview: Exploring peer review dynamics with llm agents. *arXiv preprint arXiv:2406.12708*.

William H Kruskal and W Allen Wallis. 1952. Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47(260):583–621.

Giuseppe Russo Latona, Manoel Horta Ribeiro, Tim R. Davidson, Veniamin Veselovsky, and Robert West. 2024. The ai review lottery: Widespread ai-assisted peer reviews boost paper scores and acceptance rates. *Preprint*, arXiv:2405.02150.

Ruochi Li, Haoxuan Zhang, Edward Gehringer, Ting Xiao, Junhua Ding, and Haihua Chen. 2025. Unveiling the merits and defects of LLMs in automatic review generation for scientific papers. In *2025 IEEE International Conference on Data Mining (ICDM)*, pages 1370–1379. IEEE.

Shiyang Li, Jianshu Chen, Yelong Shen, Zhiyu Chen, Xinlu Zhang, Zekun Li, Hong Wang, Jing Qian, Baolin Peng, Yi Mao, and 1 others. 2022. Explanations from large language models make small reasoners better. *arXiv preprint arXiv:2210.06726*.

Weixin Liang, Yuhui Zhang, Hancheng Cao, Binglu Wang, Daisy Ding, Xinyu Yang, Karthik Vodrahalli, Siyu He, Daniel Smith, Yian Yin, and 1 others. 2024. Can large language models provide useful feedback on research papers? a large-scale empirical analysis. *NEJM AI*, 1(8):AIoa2400196.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. 2024. The AI scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*.

Kai Lu, Shixiong Xu, Jinqiu Li, Kun Ding, and Gaofeng Meng. 2025. Agent reviewers: Domain-specific multimodal agents with shared memory for paper review. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*.

Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. 2023. Teaching small language models to reason. *arXiv preprint arXiv:2212.08410*.

Henry B Mann and Donald R Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1):50–60.

Peter McCullagh. 1980. Regression models for ordinal data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 42(2):109–127.

Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua. 2016. Ordinal regression with multiple output CNN for age estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4920–4928.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

OpenAI. 2024. Openai o1 system card. *OpenAI White Paper*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, and 1 others. 2023. Discovering language model behaviors with model-written evaluations. In *Findings of the Association for Computational Linguistics: ACL 2023*.

Qwen Team. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

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. *Advances in neural information processing systems*, 36:53728–53741.

Sascha Rothe, Jonathan Mallinson, Eric Malmi, Sebastian Krause, and Aliaksei Severyn. 2021. A simple recipe for multilingual grammatical error correction. *arXiv preprint arXiv:2106.03830*.

Mrinank Sharma, Meg Tong, Tomasz Tomasik, Jeremy Uesato, Andy Susnjak, Anna Roose, and 1 others. 2023. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances*

*in Neural Information Processing Systems*, 33:3008–3021.

Peiyi Wang, Lei Li, Liang Chen, Dawei Zhao, Binghuai Zhu, and Min Lin. 2023. Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926*.

Qingyun Wang, Lifu Huang, Heng Ji, Tom Hope, and 1 others. 2020. Reviewrobot: Explainable paper review generation based on knowledge synthesis. In *Proceedings of the 13th International Conference on Natural Language Generation*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837.

Jerry Wei, Da Huang, Yifeng Lu, Denny Zhou, and Quoc V Le. 2023. Simple synthetic data reduces sycophancy in large language models. *arXiv preprint arXiv:2308.03958*.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, and 1 others. 2024. Autogen: Enabling next-gen llm applications via multi-agent conversation. *arXiv preprint arXiv:2308.08155*.

Jiangui Yu, Zhen Ding, Jianing Tan, and 1 others. 2024. Automated peer reviewing in paper SEA: Standardization, evaluation, and analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10164–10184.

Weizhe Yuan, Pengfei Liu, and Graham Neubig. 2022. Can we automate scientific reviewing? *Journal of Artificial Intelligence Research*, 75:171–212.

Tony Z Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, pages 12697–12706. PMLR.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Shuyan Hao, Zhanghao Wu, Senhao Ba, Eugene Jiang, Chengpeng Yin, Kevin Lin, and 1 others. 2023a. Judging LLM-as-a-judge with MT-Bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36.

Yizhen Zheng, Huan Yee Koh, Jiaxin Ju, Anh T.N. Nguyen, Lauren T. May, Geoffrey I. Webb, and Shirui Pan. 2023b. Large language models for scientific synthesis, inference and explanation. *arXiv preprint arXiv:2310.07984*.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Liao, and 1 others. 2023. LIMA: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36.

Minjun Zhu, Yunfan Xu, Yuxin Wang, Zike Chen, Xintao Ren, Zhiwei Lian, Rui Wang, and Wei Yang. 2025. Deepreview: Improving llm-based paper review with human-like deep thinking process. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*. ArXiv:2503.08569.

# Appendix of "ScienceCritic: Empowering AI Review Systems with Non-objective Scientific Value Judgment"

## A Notations

Table 3 summarizes the core mathematical symbols and notations used throughout the paper.

## B Detailed Experimental Setup

### B.1 Datasets and Splitting Strategy

We evaluate on SCICRITIC-8K, a curated dataset of 8,696 manuscripts and consensus metadata spanning ICML, NeurIPS, and ICLR (2021–2025). To counteract extreme class imbalance and prevent evaluation metrics from being dominated by majority classes, we construct a strictly stratified test set: 200 Rejects, 200 Posters, 200 Spotlights, 200 Orals, and 20 Award papers, yielding a balanced suite of 820 manuscripts. The remaining 7,876 papers are allocated for training.

### B.2 Baselines

General-purpose reasoning models include **OpenAI o1** (OpenAI, 2024), a reasoning-optimized model for complex STEM tasks; **GPT-4o** (OpenAI, 2023) for academic document analysis; **Claude 3.5 Sonnet** (Anthropic, 2024) for long-context processing; **DeepSeek-V3** (DeepSeek-AI, 2024), an open-weight MoE architecture serving as the teacher; and the **Qwen2.5** family (Qwen Team, 2024) (**72B**, **32B**, and the **7B** base model). Scaling the Qwen family across parameter capacities facilitates investigation into whether model size resolves scientific judgment distributional biases. All general-purpose models are evaluated under **Zero-shot** prompting, **Few-shot** in-context learning (2 dynamic anchors), **Agentic workflows** with self-reflection, and **Agentic + Few-shot** configurations.

Specialized academic evaluation benchmarks include **Reviewer2** (Gao et al., 2024) for dynamic prompt engineering, **AgentReview** (Jin et al., 2024) for autonomous manuscript critique scaffolding, **DeepReview** (Zhu et al., 2025) for dedicated academic assessment, and **TreeReview** (Chang et al., 2025) for hierarchical reasoning structures. Detailed configurations and prompts are in Appendix G.

### B.3 Implementation Details

We instantiate SCIENCE CRITIC from Qwen2.5-7B-Instruct (Qwen Team, 2024) as the foundational base model. The model undergoes parameter-efficient Supervised Fine-Tuning (SFT) regulated by our proposed Progressive Alignment Curriculum and triple-objective optimization. Comprehensive training configurations are in Appendix F.

### B.4 Evaluation Metrics

Performance assessment utilizes ordinal and categorical metrics for value estimation $\mathbf{s}$ and final adjudication $z$. Spearman rank correlation $r_s = \rho(R(\hat{\mathbf{s}}), R(\mathbf{s}^*))$ measures monotonic alignment between predicted values $\hat{\mathbf{s}}$ and consensus labels $\mathbf{s}^* \in [1,5]^4$ via rank variable correlation. Adjudication evaluation employs overall Accuracy $Acc = \frac{1}{N}\sum \mathbb{I}(\hat{z}_i = z_i^*)$ and Per-Tier Recall $Rec_k = \frac{\sum \mathbb{I}(\hat{z}_i = k \cap z_i^* = k)}{\sum \mathbb{I}(z_i^* = k)}$ across five ordered classes $k \in \mathcal{Z}_{\text{ordered}}$. Per-Tier Recall directly diagnoses baseline positivity and central-tendency biases: $Rec_{\text{Rej}}$ isolates gatekeeping stringency while $Rec_{\text{Awd}}$ quantifies high-impact identification beyond majority-class dominance.

## C Detailed Scoring Rubric for Academic Peer Review

This appendix details the scoring rubric used for manuscript evaluation throughout this work. The four evaluation dimensions are adapted from the official review guidelines of the National Institutes of Health (NIH) and the National Science Foundation (NSF). Both the teacher model (during SCICRITIC-8K construction) and all evaluated LLMs (during inference) are prompted with these exact definitions to assign 1–5 Likert scores across all criteria.

### C.1 Evaluation Criteria Definitions

## D Constructing SCICRITIC-8K

The construction of the SCICRITIC-8K dataset relies on a two-step teacher annotation pipeline utilizing DeepSeek-V3. To ensure the reliability of our supervised fine-tuning data, we first distill unstructured expert reviews into standardized ground-truth consensus scores. Subsequently, we employ a

Table 3: Core mathematical notation used throughout the paper.

| Symbol | Type | Description |
|---|---|---|
| | | *Inputs & Dataset* |
| $x$ | Manuscript | Input scientific submission (abstract + introduction) |
| $\mathcal{D}$ | Dataset | Paired dataset of submissions and ground-truth reviews |
| $\mathcal{D}_{\text{SFT}}$ | Dataset | Distilled SFT dataset: $\{(x, \mathbf{c}^*, \mathbf{s}^*, z^*)\}$ |
| | | *Structured Output* |
| $\mathbf{s}$ | Vector | Multi-dimensional score vector; $\mathbf{s} \in [1, 5]^4$ |
| $s_1, \ldots, s_4$ | Scalar | Per-dimension scores (Strategic Scope, Gap Resolution, Intellectual Merit, Broader Impacts) |
| $\mathbf{s}^*$ | Vector | Ground-truth consensus score vector |
| $\hat{s}$ | Scalar | Model's predicted expected score: $\hat{s} = \mathbb{E}_{p_\theta}[k]$ |
| $\mathbf{c}$ | Sequence | Reasoning rationale (chain-of-thought) |
| $\mathbf{c}^*$ | Sequence | Ground-truth rationale (reasoning trajectory generation) |
| $z$ | Decision | Final acceptance tier; $z \in \mathcal{Z}_{\text{ordered}}$ |
| $z^*$, $\hat{z}$ | Decision | Ground-truth and predicted tier labels |
| $\mathcal{Z}_{\text{ordered}}$ | Ordered set | {Reject, Poster, Spotlight, Oral, Award} |
| | | *Model & Training* |
| $\theta$ | Parameters | Trainable model parameters |
| $P_\theta$ | Distribution | Generative distribution under model $\theta$ |
| $\mathcal{I}_{\text{val}}$ | Index set | Token positions of judgment values ($\mathbf{s}^*$, $z^*$) in the output sequence |
| | | *Loss Functions & Curriculum* |
| $\mathcal{L}_{\text{SFT}}$ | Loss | Standard NLL supervised fine-tuning loss |
| $\mathcal{L}_{\text{CE}}$ | Loss | Focal cross-entropy loss (amplified at judgment tokens) |
| $\mathcal{L}_{\text{Soft}}$ | Loss | Ordinal Gaussian soft-label KL-divergence loss |
| $\mathcal{L}_{\text{MSE}}$ | Loss | Expectation regression (MSE) loss |
| $\mathcal{L}_{\text{Ord}}$ | Loss | Combined ordinal objective: $\lambda_{\text{mse}}\mathcal{L}_{\text{MSE}} + \mathcal{L}_{\text{Soft}}$ |
| $q(k)$ | Distribution | Gaussian soft label: $q(k) \propto \exp\left(-\frac{(k-s^*)^2}{2\sigma^2}\right)$ |
| $\sigma$, $\lambda_{\text{mse}}$ | Scalars | Gaussian bandwidth and MSE weighting coefficient |
| $\tau_{\text{ce}}$, $\tau_{\text{acc}}$, $\tau_{\text{mse}}$ | Thresholds | Curriculum unlocking thresholds for CE, score accuracy, and MSE |
| | | *Evaluation Metrics* |
| $r_s$ | Metric | Spearman rank correlation: $\rho(R(\hat{\mathbf{s}}), R(\mathbf{s}^*))$ |
| $Acc$ | Metric | Tier-adjudication accuracy: $\frac{1}{N}\sum \mathbb{I}(\hat{z}_i = z_i^*)$ |
| $Rec_k$ | Metric | Per-tier recall for class $k$; $Rec_{\text{Rej}}$ and $Rec_{\text{Awd}}$ highlight tail-end performance |

score-conditioned reasoning trajectory generation process to generate the final analytical reasoning chains based exclusively on the manuscript text.

### D.1 Step 1: Ground-Truth Consensus Scoring

Raw peer reviews are often unstructured and qualitatively diverse. To establish a standardized target vector $\mathbf{s}^* \in [1, 5]^4$ for supervised learning, we employ the teacher model to infer the latent consensus scores from the official Meta-Review and individual reviewer comments. As detailed in Algorithm 2, this process uses a highly deterministic decoding setting ($\tau = 0.1$) to ensure systematic extraction based strictly on the predefined NIH/NSF rubrics (Appendix C).

### D.2 Step 2: Reasoning Trajectory Generation

While Algorithm 2 successfully identifies the consensus scores $\mathbf{s}^*$, the rationales generated therein are derived from reading comments of reviewers, not the manuscript itself. Training the SciCritic student model on such rationales would cause severe information leakage and hallucination during actual inference.

To resolve this, we execute a second subroutine: Score-Conditioned Reasoning Trajectory Generation (Algorithm 3). In this step, the teacher model is provided with the raw manuscript $x_{\text{student}}$ and the established ground-truth scores $\mathbf{s}^*$. It is explicitly instructed *not to predict* the score, but to read the full manuscript and extract precise methodological or empirical evidence that uniquely justifies why the manuscript deserves the target score $\mathbf{s}^*$.

**NIH & NSF Evaluation Criteria Rubrics (System Prompt Component)**

**Strategic Scope & Importance (NIH)**
The degree to which the research topic aligns with high-priority scientific challenges and opportunities, assessing the urgency and relevance of the problem context.

- **5 (Grand Challenge):** Addresses a top-tier, urgent bottleneck stalling the entire field.
- **4 (High Priority):** Addresses a timely challenge; part of a current "hot" research trend.
- **3 (Moderate):** A standard research topic; aligns with general community interests.
- **2 (Low Priority):** The topic is relevant but niche; not a current focus of the community.
- **1 (Trivial):** The topic is outdated or addresses a problem of little consequence.

**Gap Resolution & Advancement (NIH)**
The extent to which the work successfully fills a critical knowledge gap, solves a blocking problem, or introduces a valuable conceptual or technical advancement.

- **5 (Breakthrough):** Provides a complete solution to a long-standing gap; transformative.
- **4 (Significant):** Substantially advances understanding or technical capability in a clear way.
- **3 (Incremental):** Provides a modest advancement; builds predictably on existing work.
- **2 (Minimal):** Offers very limited new insight; the gap addressed is minor.
- **1 (None):** Re-evaluates known facts without adding any new resolution or advancement.

**Intellectual Merit (NSF)**
The potential of the work to advance fundamental knowledge and understanding within its own field or across different fields (theoretical contribution).

- **5 (Transformative):** Likely to change the fundamental understanding of the field or create new sub-fields.
- **4 (Extensive):** Strong potential to advance theoretical frameworks across multiple disciplines.
- **3 (Solid):** Contributes clearly to the body of knowledge within a specific domain.
- **2 (Limited):** Theoretical contribution is narrow or only provides slight variations of existing models.
- **1 (Weak):** Fails to provide theoretical insight or lacks rigorous conceptual grounding.

**Broader Impacts (NSF)**
The potential of the research to benefit society and contribute to the achievement of specific, desired societal outcomes beyond the academic community.

- **5 (High Impact):** Directly addresses a critical global crisis with high deployment potential.
- **4 (Clear Benefit):** Explicitly addresses a specific societal need with a plausible path to impact.
- **3 (Potential):** Indirect benefits exist, but specific applications are not detailed.
- **2 (Vague):** Mentions societal impact only as a buzzword without specific logic.
- **1 (None):** No foreseeable societal benefit or potential for harmful misuse.

# E Dataset Quality, Composition, and Statistical Validation

In Section 4.2, we established the global monotonic stratification of the SCICRITIC-8K dataset. This appendix provides statistical evidence for our scoring schema. Figure 6 presents the complete Mann-Whitney U test results, plotting $-\log_{10} p$ to visualize the magnitude of separation. The heatmaps confirm that statistically significant differences ($p \ll 0.001$) exist across all tier combinations, including adjacent transitions such as Spotlight versus Oral. The darker colors away from the diagonal confirm structural integrity at every quality boundary. Table 4 provides the full summary of mean and standard deviation of consensus scores for each tier.

Table 4: Mean and std of consensus scores across decision tiers.

| Tier | Scope | Gap Res. | Merit | Impacts |
|---|---|---|---|---|
| Award | $4.27 \pm 0.74$ | $4.31 \pm 0.66$ | $4.38 \pm 0.66$ | $4.30 \pm 0.71$ |
| Oral | $3.98 \pm 0.54$ | $3.89 \pm 0.65$ | $3.84 \pm 0.67$ | $3.79 \pm 0.65$ |
| Spotlight | $3.44 \pm 0.75$ | $3.46 \pm 0.73$ | $3.46 \pm 0.76$ | $3.43 \pm 0.74$ |
| Poster | $2.84 \pm 0.77$ | $2.83 \pm 0.77$ | $2.83 \pm 0.77$ | $2.84 \pm 0.77$ |
| Reject | $1.86 \pm 0.72$ | $1.89 \pm 0.72$ | $1.86 \pm 0.72$ | $1.83 \pm 0.70$ |

To provide a detailed view of the variance within each tier, Figure 7 illustrates the separate violin


Figure 6: Heatmaps of negative logarithmic $p$-values ($-\log_{10} p$) for pairwise Mann-Whitney U tests. Darker colors indicate stronger statistical separation.

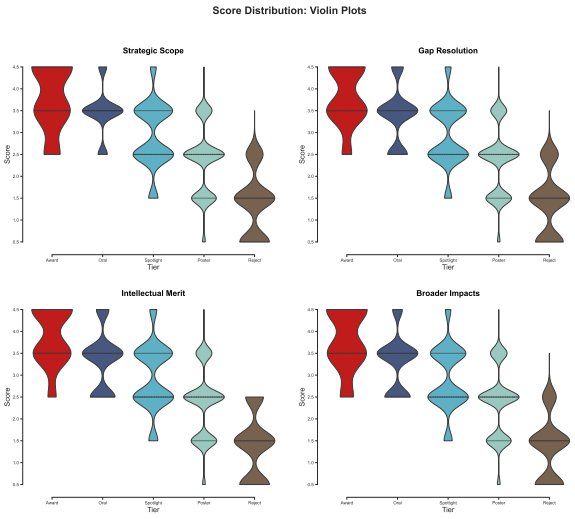


Figure 7: Detailed violin plots for each evaluation dimension. The tightening of score variance at upper tiers corresponds to stronger consensus.

**Algorithm 2** Consensus Score Extraction (Step 1)

**Input:** Manuscript Title, Official Meta-Review, Concatenated Individual Reviews.
**Output:** Ground-truth consensus scores $\mathbf{s}^*$ in a structured JSON object.
**Require:** Pre-defined Scoring Rubric $\mathcal{R}_{\text{scale}}$.
*// Phase 1: Context Initialization*
$\mathcal{P}_{\text{sys}} \leftarrow$ "Task: Evaluate scientific value by inferring consensus scores (1-5) based on the reviewers' comments and the scoring rubric $\mathcal{R}_{\text{scale}}$."
$\mathcal{P}_{\text{user}} \leftarrow \text{Concatenate}(\text{Title}, \text{Meta-Review}, \text{Reviews})$
*// Phase 2: Deterministic JSON Generation*
**Set** Decoding parameters: Temperature $\tau = 0.1$, Format = `json_object`.
$\mathcal{O}_{\text{score}} \leftarrow \text{LLM\_Generate}(\mathcal{P}_{\text{sys}}, \mathcal{P}_{\text{user}})$
*// The output enforces a strict JSON schema containing "reasoning" and "scores" for:*
*// 1. strategic_scope*
*// 2. gap_resolution*
*// 3. intellectual_merit*
*// 4. broader_impacts*
*// Phase 3: Validation and Error Handling*
**if** $\mathcal{O}_{\text{score}}$ fails JSON decoding **or** scores $\notin [1, 5]$ **then**
    $\mathcal{O}_{\text{score}} \leftarrow \text{Retry\_Generation}(\text{up to 3 times})$
**end if**
**return** $\mathbf{s}^* \leftarrow \mathcal{O}_{\text{score}}[\text{"scores"}]$

plots for each evaluation dimension. The quartile intervals (indicated by the inner box plots) clearly separate across tiers, further confirming that higher-tier papers are tightly anchored to higher, more consistent consensus scores.

## F Training Details and Hyperparameters

We summarize the hardware, optimization parameters, and curriculum configurations utilized during the Supervised Fine-Tuning (SFT) phase in Table 5.

**Hardware and Efficiency.** All models are trained on NVIDIA A100 (80GB) GPUs. To maximize computational efficiency and accommodate lengthy academic manuscripts, we enable Flash Attention 2 alongside gradient checkpointing and BF16 mixed-precision training. To handle class imbalance, we apply a Square Root sampling strategy ($\alpha = 0.5$) during batch construction, which upsamples minority classes (e.g., Award, Reject) without downsampling the majority classes.

**Progressive Alignment Curriculum Unlocking.** As detailed in Section 3.3.3, the progression of our Progressive Alignment Curriculum is strictly regulated by empirical metrics evaluated at runtime:

- **Phase 2 Unlocking (Score Metric):** The hybrid objective ($\mathcal{L}_{\text{Soft}} + \mathcal{L}_{\text{MSE}}$) for intermediate scores is unlocked only when the average weighted $\mathcal{L}_{\text{CE}}$ drops below the stability threshold of $\tau_{\text{ce}} = 1.2$.
- **Phase 3 Unlocking (Tier Decision):** The rigorous numerical constraints for the final tier adjudication token are activated only after

**Algorithm 3** Score-Conditioned Reasoning Trajectory Generation (Teacher Pipeline - Step 2)

**Input:** Raw manuscript sequence $\mathcal{X}$, Consensus vector $\mathbf{s}^* \in [1,5]^4$, Tier $z^*$.
**Output:** Verified dataset tuple $(\mathcal{X}, \mathbf{c}^*, \mathbf{s}^*, z^*)$ or **NULL**.
**Require:** Pre-defined Scoring Rubrics $\mathcal{R}_{\text{scale}}$.
*// Phase 1: Context Preparation & Hard Constraints*
**if** Length$(\mathcal{X}) > L_{\max}$ **then**
  $\tilde{\mathcal{X}} \leftarrow \mathcal{X}[1 : L_{\max}] \oplus$ "...(truncated)" {Prevent context length overflow}
**else**
  $\tilde{\mathcal{X}} \leftarrow \mathcal{X}$
**end if**
$\mathcal{P}_{\text{sys}} \leftarrow$ "**Do NOT predict scores.** Justify the provided ground truth $\mathbf{s}^*$ using concrete evidence extracted from the manuscript."
$\mathcal{P}_{\text{user}} \leftarrow \text{FormatPrompt}(\tilde{\mathcal{X}}, \mathcal{R}_{\text{scale}}, \mathbf{s}^*, z^*)$
*// Phase 2: Robust Generation Loop with Schema Validation*
$attempt \leftarrow 0$, $is\_valid \leftarrow$ False
**while** $attempt < 3$ **and not** $is\_valid$ **do**
  $attempt \leftarrow attempt + 1$
  $\mathcal{O}_{\text{raw}} \leftarrow \text{LLM\_Generate}(\mathcal{P}_{\text{sys}}, \mathcal{P}_{\text{user}}; \tau = 0.1, \text{mode} = \text{JSON})$
  *// Phase 3: Structural Parsing & Grounding Check*
  **if** TryParseJSON$(\mathcal{O}_{\text{raw}}) \rightarrow \mathcal{O}_{\text{json}}$ **then**
    $req\_keys \leftarrow$ {"strategic_scope", "gap_resolution", "intellectual_merit", "broader_impacts"}
    **if** $req\_keys \subset \text{Keys}(\mathcal{O}_{\text{json}})$ **then**
      *// Verify if the rationale strictly justifies* $\mathbf{s}^*$ *without formulaic hallucination*
      $ev\_score \leftarrow \text{CheckEvidenceGrounding}(\mathcal{O}_{\text{json}}, \tilde{\mathcal{X}})$
      **if** $ev\_score \geq \tau_{ev}$ **then**
        $\mathbf{c}^* \leftarrow \mathcal{O}_{\text{json}}$
        $is\_valid \leftarrow$ True
      **end if**
    **end if**
  **end if**
  **if not** $is\_valid$ **then**
    Sleep(ExponentialBackoff$(attempt)$) {Handle API rate limits or schema failures}
  **end if**
**end while**
*// Phase 4: Dataset Purity Enforcement*
**if** $is\_valid$ **then**
  **return** $(\mathcal{X}, \mathbf{c}^*, \mathbf{s}^*, z^*)$ {Persist high-quality reasoning to $\mathcal{D}_{\text{SFT}}$}
**else**
  **return NULL** {Discard sample to prevent supervision collapse}
**end if**

the model demonstrates foundational evaluation competence, achieving a score accuracy $\tau_{\text{acc}} \geq 0.5$.

- **Refinement Phase:** To ensure precise numerical convergence, the soft-label objective is globally disabled once the weighted expectation regression loss falls below $\tau_{\text{mse}} = 0.5$, forcing the model to exclusively optimize for exact categorical correctness.

## G Baseline Inference Configurations

We evaluate general-purpose LLMs under four distinct inference configurations (**Zero-shot**, **Few-shot**, **Agentic**, and **Agentic + Few-shot**). The implementation details for the Prompting and Workflows are described below.

**Zero-shot Prompting** The baseline model processes the full-text manuscript alongside our structured scoring rubric in a single pass. It is instructed

Table 5: Comprehensive hyperparameters and configurations for the Supervised Fine-Tuning (SFT) phase of SCIENCE CRITIC.

| **Hyperparameter** | **Value** |
|---|---|
| *Model & Architecture* | |
| Base Model | Qwen2.5-7B-Instruct |
| Max Sequence Length | 8,192 |
| Precision | BF16 |
| Gradient Checkpointing | True |
| Flash Attention 2 | Enabled |
| *LoRA Configuration* | |
| LoRA Rank ($r$) | 64 |
| LoRA Alpha ($\alpha$) | 128 |
| LoRA Dropout | 0.05 |
| Target Modules | `q, k, v, o, gate, up, down_proj` |
| *Optimization* | |
| Optimizer | AdamW |
| Learning Rate | 2e-5 |
| LR Scheduler | Cosine |
| Warmup Ratio | 0.1 |
| Weight Decay | 0.01 |
| Training Epochs | 3 |
| Per-Device Batch Size | 1 |
| Gradient Accumulation Steps | 8 |
| Effective Batch Size | 8 |
| *Triple-Objective & Curriculum Settings* | |
| Score Token Loss Weight | 20.0 |
| Tier Token Loss Weight | 30.0 |
| Ordinal Gaussian Sigma ($\sigma$) | 1.0 |
| Expectation MSE Weight | 0.5 |
| CE Stability Threshold ($\tau_{\text{ce}}$) | 1.2 |
| Score Accuracy Threshold ($\tau_{\text{acc}}$) | 0.5 |
| MSE Convergence Threshold ($\tau_{\text{mse}}$) | 0.5 |

to directly output a structured JSON object containing qualitative reasoning for each dimension, the corresponding discrete integer scores $(1-5)$, and the final tier adjudication.

**Few-shot In-Context Learning** To provide the baseline wrapper with calibration anchors, the Few-shot configuration explicitly implements a dynamic retrieval-based prompt augmentation. Before evaluating the target manuscript, the script samples two reference examples (anchors) from our curated *Source* training dataset:

- **High-Quality Anchor:** Randomly sampled from papers assigned to the positive tiers (Award, Oral, or Spotlight).
- **Low-Quality Anchor:** Randomly sampled from papers assigned to the Reject tier.

To avoid context window exhaustion, each anchor exposes only the first 3,000 characters of the source paper, concatenated with its ground-truth dimension scores and final tier label. These anchors explicitly define the boundaries of our evaluation space via in-context learning.

**Agentic Workflows** The Agentic configuration replaces the single-pass generation with a structured two-stage pipeline:

1. **Stage 1 (Memory Building):** The model reads the source manuscript and synthesizes a concise, bullet-point memory representation. This analytical memory strictly mandates the identification of key methodological claims, the enumeration of specific strengths and weaknesses across all four rubric dimensions, and preliminary draft scores ($1-5$) with a one-line justification.

2. **Stage 2 (Final Adjudication):** The context is cleared of all intermediate analytical tokens except for the structured memory generated in Stage 1. Conditioned solely on this synthesized memory and the original manuscript, the model executes its final multi-dimensional critique and outputs the structured JSON evaluation.

**Agentic + Few-shot** This configuration aggregates both methodologies by injecting the High/Low Anchor Examples into the final prompt of the Agentic 2-Stage pipeline, establishing the most competitive generalist baseline achievable through prompt engineering and workflow scaffolding constraint.

### G.1 Distributional Bias Statistics

Table 6 presents quantitative evidence for the distributional biases of off-the-shelf LLMs across all evaluation dimensions (discussed in Section 4.3).

The table is divided into three panels: Panel A shows the global standard deviation collapse of baselines compared to ground truth; Panel B shows the near-zero rate of low scores ($\leq 2.0$) assigned by baselines; and Panel C contrasts mean scores given to Reject versus Award manuscripts, illustrating the baselines' loss of discriminative resolution.

### G.2 Universal Intra-Model Analysis

Figure 9 presents the intra-model performance breakdown across representative LLM families under Zero-shot, Few-shot, and Agentic settings. The top row illustrates value estimation ($r_s$), while the bottom row visualizes Per-Tier Recall. Although in-context examples cause minor distribution shifts, all evaluated model families retain a strong central tendency bias and fail to reliably isolate distribution tails (Reject and Award). This indicates that these core behavioral issues cannot be resolved through prompt engineering or agentic workflows alone.

### G.3 Detail Error Analysis

We visualize the evaluation results using confusion matrices. To ensure the most rigorous comparison, we select the absolute best-performing baseline configuration across all evaluated models and modes (e.g., OpenAI o1 Agentic) and contrast it directly with SCIENCE CRITIC. Figure 10 presents this comparison across two dimensions: the value estimation (averaged scores $1-5$) and the final adjudication (Tiers).

The left column of Figure 10 illustrates the confusion matrices for dimensional scoring. The best-performing baseline exhibits a stark off-diagonal dispersion, fundamentally driven by the central tendency bias. Its predictions are heavily clustered around the median scores of 3 and 4, regardless of the ground-truth consensus. It explicitly avoids assigning extreme scores (1 or 5), resulting in dark vertical bands in the middle columns and near-empty regions at the tails. In contrast, SCIENCE CRITIC aligns well along the diagonal. By accurately restoring the natural variance of the score distribution, our model proves capable of penalizing fundamental flaws (score 1) and recognizing profound innovations (score 5) with high fidelity.

The right column of Figure 10 maps the final tier adjudication. The baseline's confusion matrix highlights its limitations in strict tier separation. A significant portion of ground-truth "Reject" manuscripts are misclassified into "Poster" or "Spotlight" categories, demonstrating an ingrained sycophantic positivity. Furthermore, true "Award" papers are frequently downgraded to intermediate tiers, indicating an inability to confidently isolate top-tier contributions. SCIENCE CRITIC manages to separate adjacent decision tiers clearly. The pronounced diagonal in our adjudication matrix confirms that our Progressive Alignment Curriculum and ordinal fine-tuning successfully enforce both the stringency required for rejection and the acuity needed for excellence identification.

## H Generation Order Evaluative

Chain-of-Thought (CoT) prompting—generating analytical reasoning before the final answer—is widely assumed to uniformly improve performance by emulating deliberative reasoning. Interestingly, our evaluation shows that in the domain of specialized evaluative models, direct score generation (Score-First) yields better calibration for our rigorously fine-tuned SCIENCE CRITIC compared to

Table 6: Quantitative analysis of baseline LLM distributional biases. Panel A: standard deviation. Panel B: extreme low-score ratio. Panel C: mean score by tier.

| Dimension | Ground Truth | Qwen2.5-7B | Qwen2.5-72B | DeepSeek-V3 |
|---|---|---|---|---|
| **Panel A: Global Standard Deviation** | | | | |
| Strategic Scope | 0.993 | 0.156 | 0.092 | 0.318 |
| Gap Resolution | 1.056 | 0.212 | 0.412 | 0.472 |
| Intellectual Merit | 0.985 | 0.279 | 0.492 | 0.474 |
| Broader Impacts | 1.028 | 0.449 | 0.479 | 0.592 |
| **Panel B: Extreme Low Score Ratio** ($\leq 2.0$) | | | | |
| Strategic Scope | 26.46% | 0.00% | 0.00% | 0.12% |
| Gap Resolution | 29.63% | 0.00% | 0.00% | 0.37% |
| Intellectual Merit | 26.71% | 0.00% | 0.00% | 0.24% |
| Broader Impacts | 27.07% | 2.44% | 0.00% | 9.76% |
| **Panel C: Mean Score by Tier** | | | | |
| *Reject* | | | | |
| Strategic Scope | 2.490 | 4.005 | 3.985 | 3.820 |
| Gap Resolution | 2.370 | 3.970 | 3.750 | 3.530 |
| Intellectual Merit | 2.495 | 3.910 | 3.480 | 3.185 |
| Broader Impacts | 2.410 | 3.185 | 3.385 | 3.115 |
| *Award* | | | | |
| Strategic Scope | 3.650 | 4.050 | 4.000 | 3.900 |
| Gap Resolution | 3.800 | 4.050 | 3.800 | 3.850 |
| Intellectual Merit | 3.800 | 3.900 | 3.650 | 3.500 |
| Broader Impacts | 4.500 | 3.100 | 3.200 | 3.150 |

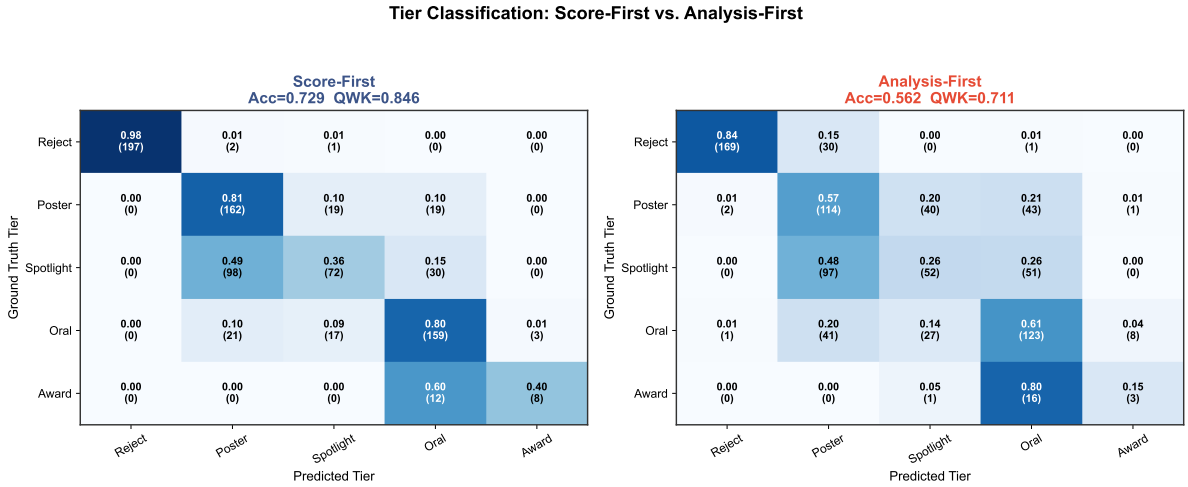


Figure 8: Confusion matrices for overall Tier adjudication ($z$). The Score-First direct latent projection ($\Phi_{\text{Sys1}}$) yields significantly sharper calibration along the diagonal ($Acc = 72.9\%$) compared to the more displaced and entropic predictions of the Rationale-First composite generative mapping ($Acc = 56.2\%$). In critically evaluating edge distributions, Score-First drastically improves gatekeeping stringency ($Rec_{\text{Rej}}$) and high-impact contribution identification ($Rec_{\text{Awd}}$).

generating rationales first (Rationale-First).

Evaluation metrics follow the definitions in Appendix B: Spearman correlation $r_s$ for value estimation and Per-Tier Recall $Rec_k$ plus overall Ac-

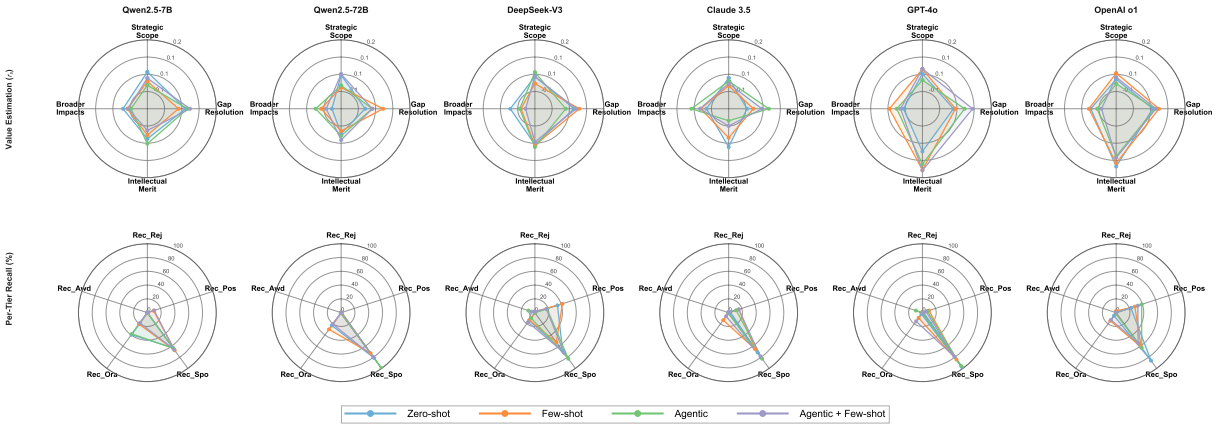


Figure 9: Intra-model performance breakdown across representative LLM families under Zero-shot, Few-shot, and Agentic settings. Top row: value estimation ($r_s$). Bottom row: Per-Tier Recall. In-context scaffolding fails to correct central tendency bias or tail-end blindness.

curacy $Acc$ for tier adjudication.

### H.1 Formalizing the Cognitive Modalities

Let $x \in \mathcal{X}$ denote the input manuscript, $\{\mathbf{s}, z\} \in \mathcal{S} \times \mathcal{Z}_{\text{ordered}}$ the multi-dimensional score vector and adjudication tier, and $\mathbf{c} = \{c_1, \ldots, c_K\} \in \mathcal{C}$ the sequence of rationale tokens. We explain the performance discrepancy by formalizing generation order as a probabilistic routing problem between two distinct inference pathways, mirroring the dual-process theory in cognitive psychology.

**System 1 (Score-First as Direct Latent Projection):** The Score-First configuration forces the model to bypass intermediate text generation, modeling a direct mapping function $\Phi_{\text{Sys1}} : \mathcal{X} \to \mathcal{S} \times \mathcal{Z}_{\text{ordered}}$. This is mathematically equivalent to predicting the objective directly from the source distribution:

$$\{\mathbf{s}, z\}_{\text{Sys1}} \sim P_{\theta^*}(\mathbf{s}, z \mid x) \tag{10}$$

Because our Triple-Objective optimization structures the model's parametric memory $\theta^*$, reading the manuscript $x$ activates relevant semantic features in the latent space. Generating scores directly reflects this representation (System 1). Without intermediate generation noise, Score-First significantly improves gatekeeping stringency ($Rec_{\text{Rej}}$) and high-impact contribution identification ($Rec_{\text{Awd}}$), yielding superior overall adjudication $Acc$ (72.9% vs. 56.2%) while consistently maximizing rank correlation $r_s$ across all value dimensions (e.g., $r_s$ for Strategic Scope improves from 0.420 to 0.694).

**System 2 (Rationale-First as Autoregressive Conditioning):** Conversely, the Rationale-First configuration explicitly unrolls its reasoning process, modeling a composite generative mapping $\Phi_{\text{Sys2}} : \mathcal{X} \to \mathcal{C} \to \mathcal{S} \times \mathcal{Z}_{\text{ordered}}$. By the probability chain rule, the final evaluation is structurally conditioned on the autoregressively sampled rationale $\mathbf{c}$ rather than exclusively on $x$:

$$\mathbf{c} \sim P_{\theta^*}(\mathbf{c} \mid x) = \prod_{k=1}^{K} P_{\theta^*}(c_k \mid x, \mathbf{c}_{<k}) \tag{11}$$

$$\{\mathbf{s}, z\}_{\text{Sys2}} \sim P_{\theta^*}(\mathbf{s}, z \mid x, \mathbf{c}) \tag{12}$$

While this sequential factorization theoretically provides a richer deliberative context window (System 2), it introduces severe cascading errors in practice.

During the generative sampling of $\mathbf{c}$, the model generates many tokens. This sequential generation introduces high variance and reasoning noise. For instance, the model might fixate on a minor syntactic issue in $x$, causing it to generate an overly critical critique sequence $\hat{\mathbf{c}}$.

Crucially, because the generated rationale $\hat{\mathbf{c}}$ is closer in the context window to the target tokens $\{\mathbf{s}, z\}$ in the decoder's KV-cache context window, the self-attention mechanism causes attention drift. The model conditions its final judgment $\{\mathbf{s}, z\}$ heavily on its own generated linguistic artifacts rather than the holistic scientific merit of the original paper $x$. The accumulation of autoregressive errors causes the predicted scores to deviate from the ground-truth consensus, leading to the lower rank-order alignment ($r_s$) and degraded accuracy ($Acc$) observed in Figure 8.

For optimal calibration in specialized evaluative models, our results strongly suggest prioritizing robust System 1 direct mapping over System 2

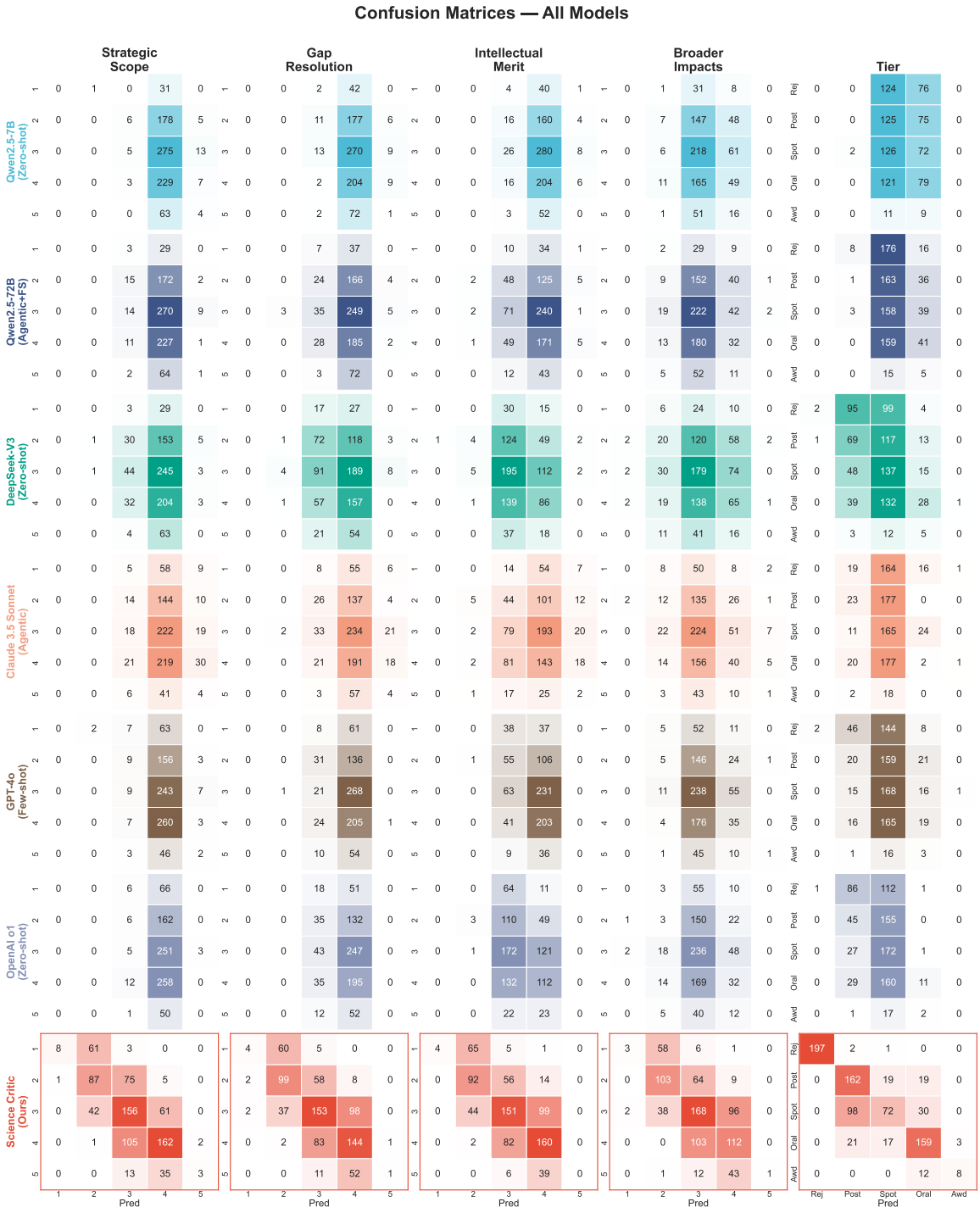


Figure 10: Confusion matrices comparing the best-performing baseline configuration (top row) against SCIENCE CRITIC (bottom row). **Left Column:** Dimensional score predictions ($1-5$). **Right Column:** Final tier adjudication (Reject to Award). The baseline matrices reveal severe central clustering and boundary blurring, whereas SCIENCE CRITIC achieves strong diagonal alignment, demonstrating strictly calibrated scoring and accurate hierarchical decision-making.

narrative conditioning. Textual rationales should primarily serve as post-hoc justifications $P_{\theta^*}(\mathbf{c} \mid x, \mathbf{s}, z)$ rather than upstream generative conditions.

# I Additional Related Work: Reasoning Distillation

Chain-of-thought (CoT) prompting (Wei et al., 2022) has emerged as a powerful technique for eliciting multi-step reasoning in large language mod-

els. A natural follow-up question is whether these reasoning abilities can be transferred from large teacher models to smaller, more efficient student models—a line of research known as CoT distillation.

**CoT Distillation.** Classical knowledge distillation (Hinton et al., 2015) transfers soft output distributions from teacher to student but does not capture intermediate reasoning processes. Recent work extends this paradigm to reasoning chains. Hsieh et al. (2023) propose *Distilling Step-by-Step*, which extracts rationales and labels jointly from a teacher model to outperform larger models with significantly less training data. Ho et al. (2023) and Magister et al. (2023) further demonstrate that CoT rationales generated by large models serve as effective supervision for smaller models, enabling them to acquire multi-step reasoning abilities. Li et al. (2022) show that LLM-generated explanations improve small model performance on reasoning benchmarks even without task-specific fine-tuning of the teacher.

**Generation Order in Reasoning.** A key design choice in CoT distillation is the ordering of reasoning and answers. Most methods adopt a *pre-thinking* paradigm where models generate rationales before answers. Chen et al. (2025) challenge this convention by proposing a *post-thinking* mechanism where the student model generates its answer before the rationale, making the answer less sensitive to minor reasoning errors. They further introduce an adaptive-thinking module that dynamically selects between pre-thinking and post-thinking based on question difficulty. This analysis of generation order resonates with our findings in Appendix H, where we observe that Score-First generation outperforms Rationale-First for our fine-tuned model.

Our Consensus Distillation pipeline (Section 3.2) differs from standard CoT distillation in that rationales are generated conditioned on established ground-truth scores $\mathbf{s}^*$, rather than being freely sampled from the teacher. This score-conditioned design prevents circular reasoning and ensures alignment with expert consensus.

## J Human Validation and Consensus Reliability

To provide empirical evidence for the reliability of the ground-truth consensus scores ($\mathbf{s}^*$) used in SciCritic-8K, we conducted a targeted human validation study involving 15 PhD and Master's students from Tsinghua University and the National University of Defense Technology.

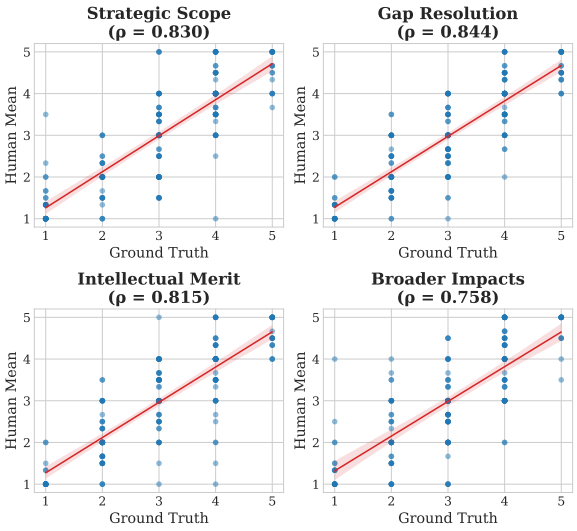


Figure 11: **Human-Ground Truth Correlation.** 2x2 grid of scatter plots illustrating the strong Spearman correlation ($\rho \in [0.758, 0.844]$) between mean human ratings and our ground-truth labels. Red lines indicate a linear regression fit (trend line), demonstrating the robust monotonic shift as scientific value increases.

The study protocol was specifically designed to mirror the consensus extraction process employed in our teacher-distillation pipeline (Section 3.2.1). Rather than performing independent manuscript assessments, human annotators were tasked with distilling the existing official expert reviews into standardized scores. For a stratified sample of 200 manuscripts, volunteers were provided solely with the official individual reviews and meta-reviews. Their goal was to project the underlying peer-review consensus onto our four-dimensional value basis (Appendix C), ensuring that the validation baseline is derived from the same deliberative evidence as the teacher-pipeline silver labels.

The quantitative results indicate strong reliability across the board. As shown in Table 7, inter-annotator agreement (Krippendorff's $\alpha$) remains in the range of 0.37–0.58. These values reflect the inherent subjectivity of mapping unstructured qualitative feedback into discrete scores, while still maintaining moderate-to-strong consensus. Furthermore, the Spearman rank correlation ($\rho$) between mean human ratings and our ground-truth labels varies from 0.758 to 0.844. As visualized in Figure 11, dimensions such as *Gap Resolution*

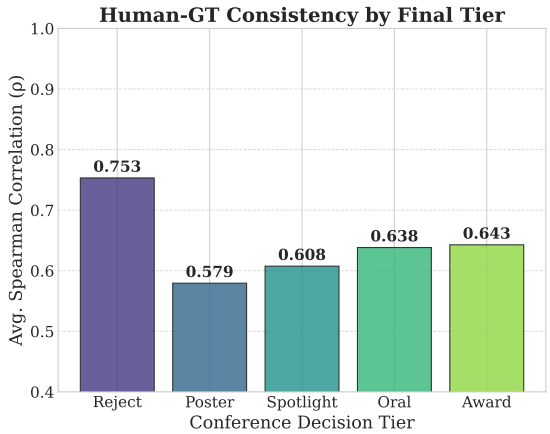


Figure 12: **Consistency across Decision Tiers.** Average Spearman correlation broken down by the final conference decision. The peak alignment in the Reject tier reflects the explicit critical signals commonly found in negative meta-reviews.

($\rho = 0.844$) and *Strategic Scope* ($\rho = 0.830$) demonstrate particularly high alignment, confirming that our consensus labels serve as a faithful proxy for expert human judgment.

Table 7: Human validation reliability metrics.

| **Dimension** | **Alpha** ($\alpha$) | **Spearman** ($\rho$) |
|---|---|---|
| Strategic Scope | 0.575 | 0.830 |
| Gap Resolution | 0.453 | 0.844 |
| Intellectual Merit | 0.546 | 0.815 |
| Broader Impacts | 0.409 | 0.758 |
| **Overall** | **0.496** | **0.812** |

Interestingly, consistency metrics vary significantly across different final decision tiers. As illustrated in Figure 12, the Spearman correlation is notably higher for manuscripts in the "Reject" category compared to accepted tiers (e.g., Spotlight or Oral). This phenomenon is largely attributed to decision leakage within the review texts. In many negative cases, both reviewers and Area Chairs use explicit, definitive language that reveals the final outcome early in the sequence (e.g., an AC stating, "I find the methodological flaws irrecoverable and will therefore recommend rejection"). Such emphatic signals reduce interpretive variance among volunteers, leading to the highly calibrated performance at the lower tail of the scientific quality distribution. Conversely, high-tier papers (Award/Oral) often involve more nuanced, subjective debates about long-term impact, slightly lowering the consensus threshold.