Mitigating Rubric Interference in LLM Judges via On-Policy Self-Distillation

Anonymous ACL submission

Abstract

LLM judges increasingly evaluate responses against fine-grained rubric checklists. When a sample requires multiple rubrics, current methods typically assess each in a separate inference call. Evaluating all rubrics in a single pass is a natural alternative with greater efficiency, but we find that it introduces *rubric interference*: the verdict on one rubric shifts depending on which other rubrics are co-present. In a preliminary study, only one-third of samples receive fully consistent verdicts when evaluated under rubric sets of varying composition. We develop a measurement framework that probes interference through four controlled operations: rubric set expansion, subsetting, reordering, and noise injection. To mitigate interference without external supervision, we propose **Self-Anchored Rubric Alignment (SARA)**. SARA uses a model's own single-rubric judgments as stable anchors and aligns multi-rubric reasoning with these anchors through on-policy self-distillation. We validate SARA on three datasets (HealthBench, FLASK, ResearchQA) and two model families (Qwen3, Llama-3.1). SARA consistently improves evaluation consistency while maintaining agreement with both base models and GPT-4.1 as a reference judge. Furthermore, the learned consistency transfers across datasets, confirming that SARA teaches a general capability rather than fitting dataset-specific patterns. Our code is available at https://anonymous.4open.science/r/SARA-22E2/

1 Introduction

Large language models now serve as automated evaluators for open-ended text generation, replacing or supplementing human annotation in both benchmarking and reward modeling (Zheng et al., 2023; Kim et al., 2024). Early approaches rely on holistic scoring or pairwise comparison. Holistic scoring collapses multidimensional quality into a single number, sacrificing diagnostic granularity (Chiang and Lee, 2023). Pairwise comparison avoids calibration issues but still provides no per-dimension feedback. These limitations have driven the field toward rubric-based evaluation, which decomposes quality into specific, verifiable items for individual assessment (Lee et al., 2025; Kim et al., 2024; Pathak et al., 2025).

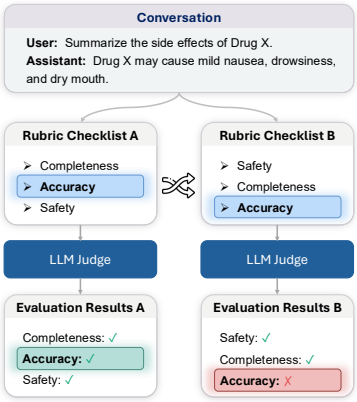

Figure 1: Rubric interference in joint evaluation. The same rubric receives different verdicts depending on the composition and ordering of co-present rubrics.

In tasks like HealthBench (Arora et al., 2025) and ResearchQA (Yifei et al., 2025), a single sample requires evaluation against multiple rubrics. Two strategies handle such multi-rubric sets. *Isolation evaluation* processes each rubric in a separate inference call, producing independent verdicts. *Joint evaluation* assesses all rubrics in a single pass, generating a unified structured report. Current practice defaults to isolation (Lee et al., 2025). Joint evaluation is attractive for its efficiency and ability to produce coherent cross-rubric narratives, both desirable when the judge serves as an online reward signal in RLHF (Gunjal et al., 2026). However, its reliability remains largely unexamined.

For joint evaluation to reliably replace isolation calls, the verdict on each rubric must depend only on the conversation and that rubric. We find that

this invariance does not hold. Joint evaluation introduces *rubric interference*: the verdict on a rubric shifts depending on how the rubric set is composed. Figure 1 illustrates one manifestation: simply reordering the rubric list flips a verdict. Interference also emerges when we expand from isolation to joint mode: comparing each rubric's isolation verdict against its joint verdict on HealthBench, even Qwen3-32B achieves only 36% sample-level exact match. This reflects systematic interference rather than generation noise: all models achieve $\geq$0.98 self-agreement across repeated evaluations with different random seeds.

Rubric interference is systematic and reproducible. It differs from known judge biases such as position, verbosity, and self-enhancement (Wang et al., 2024; Wu and Aji, 2025; Panickssery et al., 2024), which concern how a judge ranks different *responses*. Rubric interference concerns how the judge handles different *evaluation contexts* for the same response.

A key observation enables our solution. While joint evaluation is unreliable, the same model evaluating each rubric *in isolation* produces stable judgments that require no external annotation. The gap between isolation and joint verdicts reveals where interference distorts reasoning, providing a training signal without external supervision.

We propose **Self-Anchored Rubric Alignment (SARA)**, which aligns joint-mode reasoning with the model's own isolation judgments through on-policy self-distillation. SARA generates a multi-rubric evaluation, parses it into per-rubric segments, and distills from a slowly updated teacher that evaluates each segment in isolation context. A divergence loss corrects analysis segments where interference distorts reasoning; a preservation loss maintains verdict structure.

Our contributions are as follows:

1. We identify rubric interference as a systematic failure mode in multi-rubric LLM evaluation and develop a measurement framework with four operations to quantify it across binary and scored formats.
2. We propose SARA, a self-distillation method that mitigates rubric interference by anchoring joint-mode reasoning to the model's own isolation judgments, requiring no external supervision.
3. We validate SARA on three datasets and two model families. SARA improves consistency while preserving evaluation quality relative to both base models and GPT-4.1.

2 Related Work

LLM-as-a-Judge. LLM judges have become standard evaluation infrastructure (Zheng et al., 2023; Li et al., 2024). Early work showed that GPT-4 matches inter-annotator agreement on MT-Bench, justifying LLM judges at scale (Zheng et al., 2023). Chain-of-thought prompting further improves correlation with human ratings (Liu et al., 2023), and task-specific checklists from real users have moved evaluation toward structured per-sample criteria (Lin et al., 2025). More recently, rubric-based evaluation decomposes quality into fine-grained, verifiable items and shows improved reproducibility across judge models (Lee et al., 2025; Kim et al., 2024; Pathak et al., 2025). Benchmarks such as HealthBench (Arora et al., 2025) and ResearchQA (Yifei et al., 2025) adopt per-sample rubric sets as part of the evaluation protocol. Our work identifies rubric interference as a reliability concern specific to joint multi-rubric assessment.

Biases and reliability of LLM judges. LLM judges exhibit systematic biases including position bias, verbosity bias, and self-enhancement bias (Wang et al., 2024; Wu and Aji, 2025; Panickssery et al., 2024). Position bias in particular is stable and model-specific rather than a sampling artifact, persisting across 12 judge models and over 100K evaluation instances (Zheng et al., 2023). Sensitivity to prompt complexity and a general leniency tendency further compromise reliability (Thakur et al., 2025). Even GPT-4o performs near chance on challenging response pairs (Tan et al., 2025). Calibration strategies have been proposed to address these issues (Li et al., 2025, 2026). All these findings concern *inter-response* biases—how the judge compares different responses. Rubric interference is orthogonal: the judgment on a single rubric shifts depending on the co-evaluated set, with the response fixed.

Prompt sensitivity and evaluation consistency. Beyond judge-specific biases, LLMs exhibit broad prompt sensitivity: formatting changes alone cause up to 76-point accuracy swings (Sclar et al., 2024), motivating multi-prompt evaluation as a methodological default (Mizrahi et al., 2024). Paraphrase-based consistency frameworks measure whether equivalently phrased queries elicit the same factual answer (Elazar et al., 2021). These works study sensitivity of LLMs as *task-solvers*. We study sen-

sitivity of LLMs as *judges*: how context changes the label assigned to another model's output.

On-policy distillation. On-policy distillation trains a student on self-sampled trajectories with per-token teacher guidance, mitigating distribution shift (Agarwal et al., 2024; Gu et al., 2024). Self-distillation removes the need for a separate teacher: distilling a model into itself can improve generalization (Furlanello et al., 2018). In the LLM setting, models have been trained to distinguish their own outputs from references (Chen et al., 2024), to serve as their own reward function (Yuan et al., 2024), and to teach themselves by conditioning on privileged information (Zhao et al., 2026). SARA shares this on-policy self-distillation principle but targets a different goal: aligning joint-mode with isolation-mode evaluation. The privileged information is not a stronger model but the absence of interfering rubrics.

3 Problem Definition

3.1 Joint Rubric Evaluation

We consider a judge model $\mathcal{M}$ that evaluates a conversation c against a set of rubrics $R = \{r_1, \ldots, r_n\}$. For each rubric, the judge assigns a verdict v_i from a discrete value set $\mathcal{V}$. We study binary ($\mathcal{V} = \{0, 1\}$) and scored (e.g. $\mathcal{V} = \{1, \ldots, 5\}$) evaluation; the framework extends to any ordinal scale.

In **isolation evaluation**, the judge processes each rubric in a separate inference call, producing verdict v_i^{iso} from $\mathcal{M}(c, \{r_i\})$. In **joint evaluation**, the judge assesses all rubrics in a single pass, producing verdict v_i^{jnt} from $\mathcal{M}(c, R)[r_i]$. Joint evaluation is more efficient, but the verdict v_i^{jnt} may depend on the other rubrics in R, introducing rubric interference.

An ideal judge produces the same verdict for rubric r regardless of which other rubrics accompany it:

$$\mathcal{M}(c, R)[r] = \mathcal{M}(c, R')[r] \quad \forall\, R, R' \ni r \qquad (1)$$

The rest of this section describes how we test whether real judges satisfy it.

3.2 Testing for Rubric Interference

We test consistency by changing the rubric set in controlled ways and checking whether verdicts on shared rubrics remain stable. Table 1 summarizes four operations, each targeting a different interference mechanism.

Operation	What changes	What we compare
Expansion	Set size: $1 \to n$	v_i^{iso} vs v_i^{jnt}
Subsetting	Set size: $m \to n$	v_i in small vs large set
Reordering	Rubric order	v_i across permutations
Noise	Add unrelated rubrics	v_i before vs after noise

Table 1: Four operations for testing rubric interference. Each changes the rubric set while keeping the conversation and target rubric fixed.

Expansion is our primary test: we compare each rubric's isolation verdict against its joint verdict. We report *overall consistency* (full set) and *consistency-at-K* (random subsets of size K) to trace how interference grows with rubric count. **Subsetting** generalizes expansion by comparing two nested subsets of arbitrary size rather than anchoring on isolation; this tests whether interference accumulates monotonically but requires sampling valid superset. **Reordering** evaluates the same set under multiple permutations, isolating positional effects from content-based interference. **Noise injection** appends rubrics from unrelated conversations, testing whether the model can ignore irrelevant context.

For all operations, we report rubric-level agreement and Cohen's κ, plus sample-level exact match (EM). Full metric definitions are in Appendix A.

4 Self-Anchored Rubric Alignment

4.1 Overview

SARA trains a judge to resist rubric interference through on-policy self-distillation. The core idea is simple: the same model that struggles with joint evaluation judges each rubric reliably in isolation. These isolation judgments are stable and require no external annotation, making them natural training anchors.

At each training step, the student model generates a joint evaluation and parses it into per-rubric segments. A slowly updated copy of the student (the teacher) then evaluates each segment in isolation context, producing interference-free logits. A divergence loss aligns the student's joint-mode logits with these anchors on analysis segments, while a preservation loss maintains the structural tokens (verdicts, formatting). Figure 2 shows the pipeline.

4.2 On-Policy Generation and Segment Extraction

At each training step, the student $\mathcal{M}_\theta$ generates a multi-rubric evaluation for conversation c and

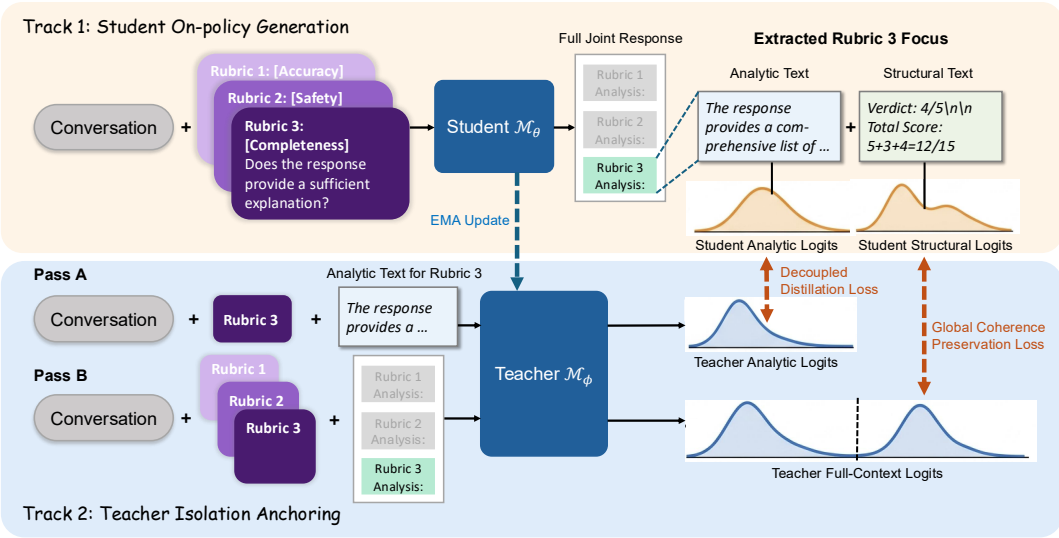

Figure 2: The SARA training pipeline. The student generates a joint evaluation (top), which is parsed into per-rubric segments. The teacher evaluates each segment under isolation context (bottom), producing interference-free anchor logits. The distillation loss aligns student and teacher logits on analysis segments; the preservation loss maintains structural tokens.

rubric set R:

$$y \sim \mathcal{M}_\theta(\cdot \mid c, R) = [a_1, s_1, \ \ldots, \ a_n, s_n] \quad (2)$$

Here a_i is the analysis segment for rubric r_i and s_i is the structural segment (verdict, formatting, transition tokens) . Because y is sampled from the current model at each step (on-policy), the training distribution tracks the model's evolving behavior.

We parse y into per-rubric segments using the structured output format (Appendix B). This produces an *analysis mask* $\mathbf{m}^{\text{ana}} \in \{0,1\}^{|y|}$ that marks tokens in all a_i segments, and a complementary *structure mask* $\mathbf{m}^{\text{str}} = \mathbf{1} - \mathbf{m}^{\text{ana}}$ that covers all s_i segments.

Segment extraction assumes instruct-mode outputs with a predictable rubric-analysis-verdict structure. Thinking-mode models require model-specific parsing, discussed in Appendix I.

4.3 Isolation-Anchored Distillation

Teacher model. The teacher $\mathcal{M}_\phi$ is a slowly updated copy of the student. After each training step:

$$\phi \leftarrow \beta_{\text{ema}}\, \phi + (1 - \beta_{\text{ema}})\, \theta \quad (3)$$

The high decay rate (β_{ema} close to 1) provides a stable reference throughout training. Using the teacher rather than the student itself as anchor avoids a degenerate loop where drifting judgments reinforce interference.

Two considerations motivate this design over using the student's own isolation logits directly. First, EMA smooths the reference temporally and prevents the training signal from oscillating at each update. Second, it breaks a potential degenerate loop: if the student anchored on its own isolation pass, any systematic bias would reinforce itself without correction.

Distillation loss. For each analysis segment a_i, we compare two forward passes over the same tokens: the student $\mathcal{M}_\theta$ under the full rubric set R, and the teacher $\mathcal{M}_\phi$ under isolation context $\{r_i\}$:

$$\begin{aligned} \mathbf{p} &= \mathcal{M}_\theta(y \mid c, \ R) \\ \hat{\mathbf{q}}^{(i)} &= \mathcal{M}_\phi(a_i \mid c, \ \{r_i\}) \end{aligned} \quad (4)$$

Here $\mathbf{p}$ denotes the student's logits over its complete joint-mode output y, conditioned on conversation c and all rubrics R. The teacher logits $\hat{\mathbf{q}}^{(i)}$ cover only the i-th analysis segment a_i, conditioned on the single rubric r_i. The student sees all rubrics and may exhibit interference; the teacher sees only the target rubric and produces interference-free logits. We align them across all analysis tokens:

$$\mathcal{L}_{\text{distill}} = \frac{1}{|\mathcal{A}|} \sum_{i=1}^{n} \sum_{t \in \mathcal{A}_i} D_{\text{JSD}}^{(\beta_d)} \left(\mathbf{p}_t \parallel \hat{\mathbf{q}}_t^{(i)} \right) \quad (5)$$

where $\mathcal{A}_i$ indexes token positions in segment a_i, $\mathcal{A} = \bigcup_i \mathcal{A}_i$ is the full analysis token set marked by $\mathbf{m}^{\text{ana}}$.

Preservation loss. Structural tokens (verdicts, formatting, transitions) need a separate signal. We

align the student's logits with the teacher's full-context logits, both conditioned on the complete rubric set R:

$$\mathbf{q} = \mathcal{M}_\phi(y \mid c,\ R) \tag{6}$$

Here $\mathbf{q}$ denotes the teacher's logits over the same output y under the same full context. The only difference from $\mathbf{p}$ is the model weights (ϕ vs θ). We compute the preservation loss over structural token positions $\mathcal{S}$ marked by $\mathbf{m}^{\text{str}}$:

$$\mathcal{L}_{\text{preserve}} = \frac{1}{|\mathcal{S}|} \sum_{t \in \mathcal{S}} D_{\text{JSD}}^{(\beta_p)}(\mathbf{p}_t \parallel \mathbf{q}_t) \tag{7}$$

We use full-context teacher logits here because structural tokens coordinate across all rubrics and have no meaningful isolation equivalent.

Total loss.

$$\mathcal{L} = \mathcal{L}_{\text{distill}} + \alpha\, \mathcal{L}_{\text{preserve}} \tag{8}$$

Since structural tokens are far fewer than analysis tokens, each structural token already receives higher per-token weight at equal α. We set α=0.4 across all experiments.

4.4 Rubric Context Augmentation

To expose the model to diverse interference patterns, the data collator applies two augmentations to each rubric set R at every training step:

- **Rubric shuffling.** We randomly permute the rubric order at each step, preventing the model from memorizing position-specific patterns.
- **Noise injection.** With probability p_{noise}, we append k rubrics from unrelated conversations, training the model to ignore irrelevant criteria.

These two augmentations target the interference sources in Table 1. Shuffling addresses positional sensitivity and noise injection addresses content-based distraction.

5 Experiments

5.1 Setup

We evaluate on three datasets spanning distinct domains and scoring formats (Table 2). HealthBench (HB) contains medical conversations evaluated against per-sample rubric checklists with binary met/unmet verdicts. Rubric counts range from 2 to 32 per sample, making it the most demanding test for interference. FLASK evaluates general-purpose instruction-following across 3 fixed skill dimensions, each scored on a 1–5 scale.

	HB	FLASK	RQA
Domain	Medical	General	Scientific
Scoring	Binary	1–5	0–4
Rubrics/sample	2–32	3 (fixed)	1–8
Avg rubrics	11.5	3.0	7.5
Samples	500	1740	1000

Table 2: Dataset statistics. HB = HealthBench, RQA = ResearchQA.

ResearchQA (RQA) evaluates scientific question answering against per-sample rubrics scored on a 0–4 scale. We sampled 1000 instances from the full dataset.

Models and baselines. We apply SARA to Qwen3-8B, Qwen3-14B, Qwen3-32B, and Llama-3.1-8B-Instruct. As a baseline, we compare against **SFT$_{\text{iso}}$**: standard supervised fine-tuning on isolation outputs concatenated into joint format. This represents the most direct approach to teaching joint-mode behavior from isolation references. Each dataset is split 8:1:1 for training, validation, and test.

Training details. SARA uses full-parameter fine-tuning with EMA decay 0.999, symmetric JSD (β_d=0.5), KL preservation (β_p=1.0, α=0.4), and vLLM greedy decoding for on-policy generation. We select checkpoints by shuffle consistency on the validation set.

Evaluation. All evaluations use greedy decoding with fixed random seeds. We cap rubrics at 10 per sample. For shuffle tests, we average over 5 random permutations. We report rubric-level agreement (Agr), Cohen's κ (weighted for scored formats), and sample-level exact match (EM).

Infrastructure. All experiments run on 8× NVIDIA H20 GPUs with vLLM for batched inference.

5.2 Main Results

Table 3 presents overall consistency between isolation and joint mode. SARA delivers consistent improvements across model scales, architectures, and evaluation domains.

The largest gains appear on ResearchQA: Qwen3-8B's EM nearly triples (.22→.59), and its κ jumps from .665 to .868. ResearchQA combines moderate rubric counts with per-sample rubric definitions, creating ample opportunity for cross-rubric contamination. On HealthBench, binary verdicts set a higher baseline and lower ceiling, yet improvements remain steady across models up to 14B. On FLASK, only 3 fixed rubrics limit the surface area

	HealthBench			ResearchQA			FLASK		
	Agr↑	κ↑	EM↑	Agr↑	κ↑	EM↑	Agr↑	κ↑	EM↑
Qwen3-8B	.819	.630	.22	.774	.665	.22	.720	.663	.41
+*SARA*	**.874**	**.736**	**.34**	**.914**	**.868**	**.59**	**.810**	**.801**	**.52**
Qwen3-14B	.874	.747	.34	.768	.658	.23	.722	.665	.41
+*SARA*	**.903**	**.806**	**.38**	**.900**	**.879**	**.52**	**.803**	**.802**	**.55**
Qwen3-32B	.887	.775	.36	.720	.651	.25	.701	.652	.41
+*SARA*	.881	.762	.36	**.880**	**.851**	**.42**	**.836**	**.822**	**.59**
Llama-3.1-8B	.733	.467	.08	.711	.608	.12	.609	.637	.24
+*SARA*	**.830**	**.652**	**.18**	**.853**	**.800**	**.34**	**.722**	**.760**	**.41**

Table 3: Overall consistency (isolation vs. joint mode). Agr = rubric-level agreement, κ = Cohen's kappa (weighted for scored formats), EM = sample-level exact match.

for interference, yet gains persist (e.g., Qwen3-32B: EM .41→.59), confirming that even minimal rubric sets trigger measurable interference.

SARA benefits weaker models more (Llama-3.1-8B: +14.2% Agr on ResearchQA), consistent with our hypothesis that stronger models have implicitly learned partial invariance through scale. Yet even Qwen3-32B improves substantially on scored formats, suggesting scale alone does not fully resolve interference. EM gains are disproportionately large relative to Agr across all settings (e.g., Qwen3-8B on ResearchQA: Agr +.140 but EM +.370). This asymmetry reveals that baseline interference is not concentrated on a few hard rubrics but scattered across many—a single flipped rubric per sample suffices to break exact match, and SARA systematically corrects these distributed errors. On scored datasets, SARA also narrows the gap between Agr and κ (Qwen3-32B on ResearchQA: .720/.651→.880/.851), indicating that remaining disagreements concentrate at adjacent scores (± 1) rather than large deviations—interference is not only less frequent but less severe.

5.3 Consistency Without Quality Degradation

	Consistency			Iso Agr w/	Jnt Agr w/
	Agr↑	κ↑	EM↑	Base*↑	Base*↑
Qwen3-14B	.768	.658	.23	–	–
+SFT_{iso}	**.909**	.872	**.54**	.892	.777
+SARA	.900	**.879**	.52	**.925**	**.806**

Table 4: SFT_{iso} vs. SARA on ResearchQA. Left: consistency between evaluation modes. Right: agreement with the untrained base model's outputs.

Why not naive SFT? A natural baseline is to directly fine-tune on isolation outputs concatenated into joint format (SFT_{iso}: 5 random-seed isolation samples per rubric as synthetic joint targets). Table 4 compares this approach against SARA on Qwen3-14B (ResearchQA). Beyond consistency, we measure behavioral drift: "Iso. Agr with Base*" reports how much a trained model's isolation outputs agree with the untrained base; "Joint Agr with Base*" does the same for joint outputs.

SFT_{iso} achieves slightly higher iso–joint consistency (Agr .909 vs .900, EM .54 vs .52), but drifts further from the original model's behavior in both modes (Iso: .892 vs .925; Joint: .777 vs .806). SFT forces the model to *reproduce* fixed target outputs, inadvertently shifting its judgment distribution; SARA teaches the model to *reason independently per rubric*, preserving its original evaluation behavior while reducing interference.

	Iso. Agr w/		Joint Agr w/	
	Base*	GPT-4.1	Base*	GPT-4.1
Qwen3-8B	–	.729	–	.773
+*SARA*	.921	.706	.874	.731
Qwen3-14B	–	.801	–	.795
+*SARA*	.912	.806	.896	.806
Qwen3-32B	–	.830	–	.815
+*SARA*	.896	.823	.894	.819
Llama-3.1-8B	–	.702	–	.726
+*SARA*	.839	.700	.755	.755

Table 5: Evaluation quality on HealthBench. Base*: agreement with untrained base. GPT-4.1: external reference judge.

Does SARA preserve evaluation quality? Table 5 confirms across all models on HealthBench that SARA's consistency gains do not come at the expense of judgment quality. Agreement with the untrained base exceeds .84 in all settings, confirming minimal behavioral drift. GPT-4.1 agreement is preserved or improved (Qwen3-14B joint: .795→.806; Llama-3.1-8B joint: .726→.755): when isolation judgments are already reasonable, reducing interference also improves absolute accuracy.

5.4 Interference Under Controlled Perturbations

We apply the perturbation framework from Table 1 to trace how each interference source affects model behavior, and how SARA mitigates it.

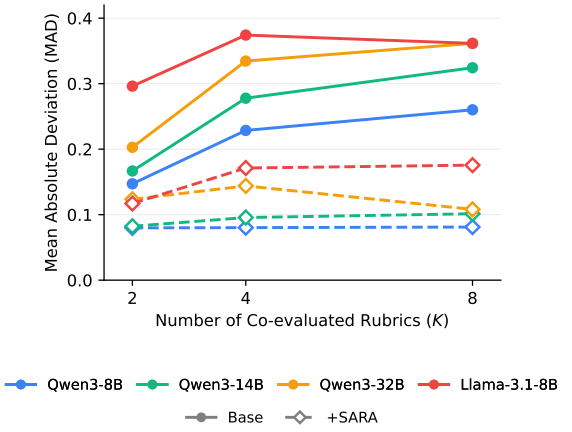

Figure 3: Mean Absolute Deviation (MAD) between isolation and joint verdicts on ResearchQA as co-evaluated rubric count K grows.

Scaling with rubric count. Figure 3 traces MAD on ResearchQA as K grows from 2 to 8. For all base models, MAD rises steeply with K: Qwen3-32B climbs from .203 at K=2 to .362 at K=8, nearly doubling. SARA flattens these curves. Qwen3-8B's MAD stays nearly constant (.080 at K=2, .080 at K=4, .081 at K=8), indicating that the model evaluates each rubric independently of set size. Across all models, SARA compresses the MAD range from .147–.374 (base) to .080–.176 (trained) and eliminates the upward trend with K.

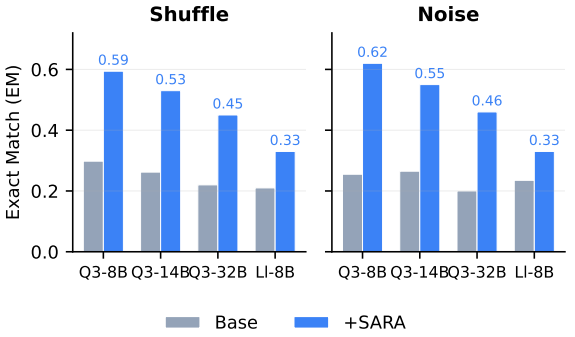

Figure 4: Shuffle invariance and noise robustness on ResearchQA (EM).

Shuffle and noise robustness. Figure 4 reports EM under random rubric permutation (shuffle) and irrelevant rubric injection (noise) on ResearchQA. Under shuffling, Qwen3-14B achieves only .262 EM; under noise, Qwen3-32B drops to .200. SARA roughly doubles shuffle EM across all models (e.g., Qwen3-8B: .298→.594) and yields similar noise gains (Qwen3-8B: .255→.620). Full results across all datasets are in Appendix D.

5.5 SARA Learns Transferable Consistency

To confirm that SARA teaches a general capability rather than fitting dataset-specific patterns, we evaluate Qwen3-14B trained on ResearchQA (SARA_{rqa}) directly on HealthBench and FLASK without further fine-tuning (Table 6).

HealthBench (Agr / κ / EM)			
Qwen3-14B	.874	.747	.34
+SARA	**.903**	**.806**	**.38**
+SARA_{rqa}	.890	.779	.34
FLASK (Agr / κ / EM)			
Qwen3-14B	.722	.665	.41
+SARA	.803	**.802**	**.55**
+SARA_{rqa}	**.820**	.800	**.55**

Table 6: Cross-dataset transfer. SARA_{rqa} (trained on ResearchQA) evaluated zero-shot on HealthBench and FLASK.

SARA_{rqa} improves over the base model on both target datasets. On FLASK it even outperforms the in-domain SARA model in Agr (.820 vs .803). This confirms that SARA learns a domain-agnostic ability to decouple per-rubric reasoning from contextual interference.

5.6 Analysis

Attention analysis. To understand how SARA reduces interference mechanistically, we compare attention patterns during joint evaluation (Qwen3-14B, ResearchQA, 10 test samples). Figure 5 shows the average attention matrix: for each output analysis segment (row), we compute what fraction of its attention flows to each input rubric span and each output analysis span. Each row sums to one.

Two effects stand out. First, SARA concentrates output self-attention on the diagonal: the average self-attention rises from .623 to .809 (+30% relative), meaning each rubric's analysis attends predominantly to its own prior reasoning rather than to other rubrics' analyses. Second, SARA suppresses cross-rubric leakage. Off-diagonal attention in the output half drops from .039 to .014 (–64% relative). The base model's lower triangle shows visible leakage—later rubrics attend heavily to earlier rubrics' analyses (e.g., Out R2 allocates .168 to Out R1). SARA nearly eliminates this pattern (the same cell drops to .048).

The input half tells a consistent story: attention to other input rubrics decreases from .018 to .011

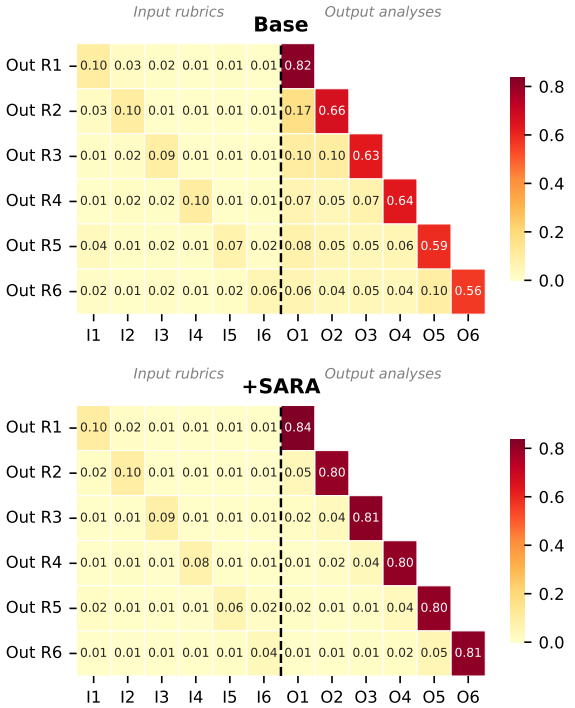

Figure 5: Normalized attention from each output analysis segment to input rubric spans (I1–I6) and output analysis spans (O1–O6), averaged over 10 samples.

(–38% relative), confirming that SARA's reasoning becomes less sensitive to the presence of co-evaluated rubric definitions. Together, these results show that SARA achieves interference reduction through a concrete mechanistic change: redirecting attention from cross-rubric spans back to the model's own analysis of the target rubric.

Case study. Table 7 shows a FLASK example where interference distorts verdicts bidirectionally. The task is a fill-in-the-blank question about business ethics; the assistant provides a plausible but imprecise answer without citations. Both models agree in isolation: R1 (comprehension) and R2 (factuality) deserve 3/5, while R3 (readability) deserves 4–5/5.

In joint mode, the base model's R1 inflates from 3 to 4 (softer phrasing: "somewhat relevant" replaces "lacks precision and completeness"), while R2 deflates from 3 to 2 (absolutist language: "does not provide *any* factual evidence" replaces "the concept *is* accurate … not supported by citation"). SARA's joint output matches isolation on all three rubrics, preserving both the critical tone of R1 and the balanced assessment of R2. Full outputs are in Appendix H.

Additional observations. Qwen3-32B shows limited gains on HealthBench because the base model already achieves .887 Agr and binary verdicts

Task: Fill in the blank about ethics management. **Response**: "...business ethics management." (no citation)
R2 (Factuality): Is the response supported by reliable evidence or citation?
Base — isolation (3/5): *"While the concept of managing ethical issues through policies and programs is accurate, the specific term 'business ethics management' is not clearly defined or supported by a citation."* **Base — joint (2/5 ✗):** *"The response does not provide any factual evidence or citation to support the claim … there is no reliable source explicitly referenced to confirm its accuracy."* **SARA — joint (3/5 ✓):** *"The response lacks any citation or reference to a reliable source … While the concept may be accurate, the support is incomplete and not fully reliable."*

Table 7: Case study (Qwen3-14B, FLASK). Base model's joint R2 adopts absolutist language, deflating 3/5 to 2/5. SARA maintains the balanced assessment.

provide a coarser distillation signal than scored verdicts. On scored datasets, gains are clear (ResearchQA: .720→.880; FLASK: .701→.836). SARA-trained models also generate longer outputs (e.g., Qwen3-8B on HealthBench: 405→611 tokens) due to more detailed per-rubric analyses that match isolation depth. Full token statistics are in Appendix G.

We report hyperparameter sensitivity and preservation loss ablation of SARA in Appendix F.

6 Conclusion

We identify rubric interference as a systematic failure mode in multi-rubric LLM evaluation and propose SARA, a self-distillation method that aligns joint-mode reasoning with the model's own isolation judgments. SARA improves consistency across datasets, model families, and scoring formats, and transfers across domains without retraining. Attention analysis confirms the mechanism: SARA selectively suppresses cross-rubric information flow while preserving within-rubric coherence.

Our results suggest a broader principle: when a model possesses a capability but fails to exercise it under complex conditions, self-distillation from a simpler setting can close the gap without external supervision. Looking ahead, SARA's training signal complements standard SFT and RLHF objectives; combining interference resistance with reward model training could yield judges that are both accurate and consistent. The measurement framework itself can also serve as a diagnostic protocol for any multi-rubric judge before deployment.

Limitations

Thinking-mode models. SARA's segment extraction relies on a predictable rubric-analysis-verdict structure in instruct-mode outputs. Adapting SARA to thinking-mode models requires paragraph-level heuristics that are less precise, and preliminary results show smaller gains than in instruct mode (Appendix I). Achieving stronger results in thinking mode likely requires additional format supervision, such as training the model to use explicit rubric delimiters within its thinking block.

Isolation as anchor. SARA treats isolation verdicts as interference-free anchors. This assumption is supported by the high self-agreement of isolation judgments ($\geq$0.98 across all models), but isolation is not infallible. In some cases, co-evaluating related rubrics may surface useful context that improves judgment quality. Our framework does not distinguish beneficial context from harmful interference. Developing methods that preserve helpful inter-rubric information while suppressing harmful interference remains an open direction.

Ethics Statement

Our work improves the consistency of LLM-based evaluation. We note that more consistent automated judges may encourage reduced human oversight; however, consistency does not guarantee correctness, and we recommend SARA-trained models complement rather than replace human evaluation in high-stakes domains. All datasets and models used are publicly available under permissive licenses, and no personally identifiable information is collected or processed.

References

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. 2024. On-policy distillation of language models: Learning from self-generated mistakes. *Preprint*, arXiv:2306.13649.

Rahul K. Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, Johannes Heidecke, and Karan Singhal. 2025. Healthbench: Evaluating large language models towards improved human health. *Preprint*, arXiv:2505.08775.

Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. 2024. Self-play fine-tuning converts weak language models to strong language models. *Preprint*, arXiv:2401.01335.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. 2021. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031.

Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. 2018. Born again neural networks. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1607–1616. PMLR.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. 2024. Minillm: Knowledge distillation of large language models. In *International Conference on Learning Representations*, volume 2024, pages 32694–32717.

Anisha Gunjal, Anthony Wang, Elaine Lau, Vaskar Nath, Yunzhong He, Bing Liu, and Sean M. Hendryx. 2026. Rubrics as rewards: Reinforcement learning beyond verifiable domains. In *The Fourteenth 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. 2024. Prometheus 2: An open source language model specialized in evaluating other language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4334–4353, Miami, Florida, USA. Association for Computational Linguistics.

Yukyung Lee, Joonghoon Kim, Jaehee Kim, Hyowon Cho, Jaewook Kang, Pilsung Kang, and Najoung Kim. 2025. Checkeval: A reliable llm-as-a-judge framework for evaluating text generation using checklists. *Preprint*, arXiv:2403.18771.

Haitao Li, Junjie Chen, Qingyao Ai, Zhumin Chu, Yujia Zhou, Qian Dong, and Yiqun Liu. 2025. CalibraEval: Calibrating prediction distribution to mitigate selection bias in LLMs-as-judges. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16537–16552, Vienna, Austria. Association for Computational Linguistics.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024. Llms-as-judges: A comprehensive survey on llm-based evaluation methods. *Preprint*, arXiv:2412.05579.

Qingquan Li, Shaoyu Dou, Kailai Shao, Chao Chen, and Haixiang Hu. 2026. Evaluating scoring bias in llm-as-a-judge. *Preprint*, arXiv:2506.22316.

Bill Yuchen Lin, Yuntian Deng, Khyathi Chandu, Abhilasha Ravichander, Valentina Pyatkin, Nouha Dziri, Ronan Le Bras, and Yejin Choi. 2025. Wildbench: Benchmarking LLMs with challenging tasks from real users in the wild. In *The Thirteenth International Conference on Learning Representations*.

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, Singapore. Association for Computational Linguistics.

Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. 2024. State of what art? a call for multi-prompt LLM evaluation. *Transactions of the Association for Computational Linguistics*, 12:933–949.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. 2024. Llm evaluators recognize and favor their own generations. *Preprint*, arXiv:2404.13076.

Aditya Pathak, Rachit Gandhi, Vaibhav Uttam, Arnav Ramamoorthy, Pratyush Ghosh, Aaryan Raj Jindal, Shreyash Verma, Aditya Mittal, Aashna Ased, Chirag Khatri, Yashwanth Nakka, Devansh, Jagat Sesh Challa, and Dhruv Kumar. 2025. Rubric is all you need: Improving llm-based code evaluation with question-specific rubrics. In *Proceedings of the 2025 ACM Conference on International Computing Education Research V.1*, ICER '25, page 181–195. ACM.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. Quantifying language models'sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *International Conference on Learning Representations*, volume 2024, pages 25055–25083.

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

Aman Singh Thakur, Kartik Choudhary, Venkat Srinik Ramayapally, Sankaran Vaidyanathan, and Dieuwke Hupkes. 2025. Judging the judges: Evaluating alignment and vulnerabilities in LLMs-as-judges. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²)*, pages 404–430, Vienna, Austria and virtual meeting. Association for Computational Linguistics.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450, Bangkok, Thailand. Association for Computational Linguistics.

Minghao Wu and Alham Fikri Aji. 2025. Style over substance: Evaluation biases for large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 297–312, Abu Dhabi, UAE. Association for Computational Linguistics.

Li S. Yifei, Allen Chang, Chaitanya Malaviya, and Mark Yatskar. 2025. Researchqa: Evaluating scholarly question answering at scale across 75 fields with survey-mined questions and rubrics. *Preprint*, arXiv:2509.00496.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason E Weston. 2024. Self-rewarding language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 57905–57923. PMLR.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. 2026. Self-distilled reasoner: On-policy self-distillation for large language models. *Preprint*, arXiv:2601.18734.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.

A Detailed Metric Definitions

All experiments compare verdicts on the same conversation–rubric pair (c, r_i) across two evaluation conditions A and B (e.g., isolation vs. joint, or two different rubric permutations). We use the same core metrics for both binary and scored formats.

A.1 Rubric-Level Metrics

Given N verdict pairs $\{(v_i^{(A)}, v_i^{(B)})\}_{i=1}^{N}$:

Agreement (Agr). Fraction of rubrics with identical verdicts:

$$\text{Agr} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}[v_i^{(A)} = v_i^{(B)}] \tag{9}$$

For binary verdicts, this is equivalent to classification accuracy between two annotators. For scored verdicts, it measures exact score match (identical to what some prior work calls Exact Agreement Rate).

Cohen's κ. Agreement corrected for chance:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{10}$$

where $p_o = \text{Agr}$ and p_e is the expected agreement under independence. For binary verdicts, we use standard (unweighted) κ. For scored verdicts, we use quadratic-weighted κ, which penalizes larger deviations more heavily. Both variants are reported under the same symbol κ throughout; the weighting scheme is determined by the scoring format.

Mean Absolute Deviation (MAD). Average per-rubric deviation:

$$\text{MAD} = \frac{1}{N} \sum_{i=1}^{N} |v_i^{(A)} - v_i^{(B)}| \tag{11}$$

For binary verdicts, MAD $= 1 -$ Agr. For scored verdicts, MAD captures the typical magnitude of disagreement.

Root Mean Squared Error (RMSE).

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (v_i^{(A)} - v_i^{(B)})^2} \tag{12}$$

More sensitive to large deviations than MAD. Reported in appendix tables for scored formats.

Pearson Correlation (ρ). Linear correlation between $v^{(A)}$ and $v^{(B)}$. Measures whether the two conditions preserve relative ordering of scores.

A.2 Sample-Level Metric

Exact Match (EM). Fraction of samples where *all* rubrics agree:

$$\text{EM} = \frac{1}{|C|} \sum_{c \in C} \mathbb{1}\left[\forall\, r_i \in R_c :\ v_i^{(A)} = v_i^{(B)}\right] \tag{13}$$

EM is the strictest metric: a single rubric disagreement within a sample counts as a failure.

A.3 Aggregate-Level Metrics

For scored formats, we also compute the total score per sample $T_c^{(X)} = \sum_{r_i \in R_c} v_i^{(X)}$ and report:

Total Exact Match (TotalEM). Fraction of samples with identical total scores:

$$\text{TotalEM} = \frac{1}{|C|} \sum_{c \in C} \mathbb{1}[T_c^{(A)} = T_c^{(B)}] \tag{14}$$

Total MAE. Mean absolute error of total scores:

$$\text{TotalMAE} = \frac{1}{|C|} \sum_{c \in C} |T_c^{(A)} - T_c^{(B)}| \tag{15}$$

Total Pearson (ρ_T). Pearson correlation of total score sequences. Measures whether sample-level ranking is preserved across conditions.

A.4 Metric Selection

Metric	HB	FLASK	RQA
Main tables (all formats)			
Agr, κ, EM	✓	✓	✓
Appendix (scored only)			
MAD, RMSE, ρ		✓	✓
TotalEM, TotalMAE, ρ_T		✓	✓
Operations			
Overall, Con-K	✓	✓	✓
Subsetting	✓		✓
Shuffle	✓	✓	✓
Noise	✓	✓	✓

Table 8: Metric and operation applicability. FLASK's fixed 3-rubric format does not support Con-K or Subsetting.

B Evaluation Prompts

We use the following task prompts for joint-mode evaluation. In isolation mode, the same prompt is used with a single-rubric list. The FLASK prompt follows the same format as ResearchQA with the scoring scale adapted to 1–5.

B.1 HealthBench (Binary)

Joint-Mode Evaluation Prompt (Binary)

```
# Task
You are a professional dialogue grader. Your
task is to evaluate the **last response
(completion)** from the Assistant in the
provided conversation against a list of
scoring rubrics. You must determine which
criteria are met, and output only the indices
of the met rubrics.
———————
# Input Data
**Conversation**:
{conversation}
———————
**Rubrics**:
{rubrics}
(Each rubric contains a "criterion" and a
"point_value". Points can be positive or
negative.)
———————
# Scoring Logic
1. **Evaluation**: For each item in the
Rubrics, determine if the Assistant's last
response meets the specific criterion.
2. **"Such as" Exception**: If a criterion
includes phrases like "such as", "for
example", or "including", the response
meets the criteria if it addresses the core
requirement, even if it does not mention every
specific example listed.
```

```
3. **Negative Criteria**: If a criterion describes an undesirable behavior (e.g., "is overly verbose" or "contains inaccuracies"), the criteria is "met" only if the Assistant actually exhibited that bad behavior.

# Output Format
For each rubric, output strictly in this format (one entry per rubric, do NOT split a single rubric into multiple entries):

**Rubric 1 Analysis**: <analysis>. **Verdict**: Met/Not Met.

**Rubric 2 Analysis**: <analysis>. **Verdict**: Met/Not Met.

...

Met Rubric Indices: [list of index numbers, separated by commas]
```

B.2 ResearchQA / FLASK (Scored)

Joint-Mode Evaluation Prompt (Scored)

```
# Task
You are a professional dialogue grader. Your task is to evaluate the **last response (completion)** from the Assistant in the provided conversation against a list of scoring rubrics. Each rubric describes a specific requirement. You must judge the degree to which the response fulfills each requirement on a 1-5 scale.

———

# Input Data
**Conversation**:
{conversation}

———

**Rubrics**:
{rubrics}
(Each rubric describes a specific requirement that the response should fulfill.)

———

# Scoring Logic
1. **Evaluation**: For each rubric, determine how well the Assistant's last response fulfills the described requirement.
2. **Scoring Scale**:
- **5**: Fully fulfilled — the response completely and accurately addresses the requirement.
- **4**: Mostly fulfilled — the response addresses the requirement with minor omissions or imprecisions.
- **3**: Partially fulfilled — the response addresses some aspects but misses key elements.
- **2**: Minimally fulfilled — the response only tangentially or superficially relates to the requirement.
- **1**: Not fulfilled — the response fails to address the requirement or directly contradicts it.
3. **Holistic Judgment**: If the response falls between two levels, choose the lower one.

# Output Format
For each rubric, output strictly in this format (one entry per rubric, do NOT split a single rubric into multiple entries):

**Rubric 1 Analysis**: <analysis>. **Verdict**: <score>/5.

**Rubric 2 Analysis**: <analysis>. **Verdict**: <score>/5.

...

Total Score: <s1>+<s2>+...=<sum>
```

The FLASK prompt uses the identical format with the scoring scale adjusted to match its 1–5 skill-based rubrics. For isolation-mode evaluation, the same prompts are used with $|R| = 1$ (a single rubric in the rubric list). The output format reduces to a single analysis–verdict pair.

C Consistency-at-K Full Results

Tables 9 and 10 report the full Consistency-at-K results underlying Figure 3. FLASK is omitted because its fixed 3-rubric format does not support variable-K evaluation.

D Shuffle Invariance and Noise Robustness Full Results

Figure 6 summarizes shuffle invariance and noise robustness results. Full numerical results are in Tables 11 and 12.

E Supplementary Metrics for Main Experiment

Table 13 reports additional rubric-level metrics (MAD, RMSE, Pearson ρ) for the scored datasets in the main experiment (overall consistency: isolation vs. joint mode). Table 14 reports aggregate-level metrics computed from per-sample total scores.

F Hyperparameter Sensitivity

Hyperparameter sensitivity. Table 15 ablates the two key hyperparameters on Qwen3-14B (ResearchQA). Setting EMA decay=1 (teacher equals student, no momentum averaging) drops EM from .52 to .48, confirming that the teacher must be decoupled from the student to provide a stable distillation target. Performance is robust across decay $\in [.99, .999]$. The JSD interpolation weight β_d shows similar robustness: EM varies by only .02 across $\beta_d \in [0.1, 0.7]$. However, β_d=0.9 causes EM to collapse to .38—when the loss is dominated by the teacher distribution, the student loses the capacity to produce coherent independent judgments.

	$K = 2$			$K = 4$			$K = 8$		
	Agr↑	κ↑	EM↑	Agr↑	κ↑	EM↑	Agr↑	κ↑	EM↑
HealthBench									
Qwen3-8B	.871	.722	.754	.844	.672	.515	.843	.675	.266
+*SARA*	.914	.816	.832	.896	.779	.654	.885	.756	.403
Qwen3-14B	.901	.802	.806	.890	.779	.635	.874	.746	.355
+*SARA*	.953	.905	.912	.932	.864	.746	.912	.864	.468
Qwen3-32B	.895	.789	.796	.882	.763	.600	.884	.768	.350
+*SARA*	.913	.824	.836	.900	.798	.654	.898	.795	.415
Llama-3.1-8B	.797	.594	.627	.749	.498	.309	.733	.469	.094
+*SARA*	.861	.718	.741	.821	.636	.467	.824	.641	.240

Table 9: Consistency-at-K on HealthBench. Metrics: Agr = rubric-level agreement, κ = Cohen's kappa, EM = sample-level exact match.

	$K = 2$			$K = 4$			$K = 8$		
	Agr↑	MAD↓	EM↑	Agr↑	MAD↓	EM↑	Agr↑	MAD↓	EM↑
ResearchQA									
Qwen3-8B	.867	.147	.761	.808	.229	.487	.787	.260	.216
+*SARA*	.928	.080	.866	.927	.080	.751	.922	.081	.568
Qwen3-14B	.867	.167	.751	.803	.278	.460	.787	.324	.243
+*SARA*	.926	.082	.862	.913	.096	.705	.909	.101	.460
Qwen3-32B	.837	.203	.718	.756	.335	.405	.760	.362	.216
+*SARA*	.892	.124	.797	.876	.144	.592	.910	.108	.378
Llama-3.1-8B	.788	.296	.628	.739	.374	.324	.740	.362	.027
+*SARA*	.889	.117	.809	.862	.171	.563	.872	.176	.324

Table 10: Consistency-at-K on ResearchQA. Agr here is exact score agreement (identical to EAR in prior work). MAD captures deviation magnitude.

The preservation loss is necessary for structural integrity. Setting $\alpha = 0$ (no preservation loss) causes the model to lose the ability to produce structural outputs within tens of training steps; even $\alpha = 0.1$ makes a difference. Unlike the EMA and β_d parameters (Table 15), the preservation coefficient is not a tunable dial but a binary structural requirement.

G Output Token Statistics

Table 16 reports the average number of generated tokens per sample in joint mode. SARA-trained models produce longer outputs due to more detailed per-rubric analyses. The increase is proportional to the average rubric count: HealthBench (avg 11.5 rubrics) shows the largest absolute increase, while FLASK (3 rubrics) shows the smallest.

H Case Study: Full Outputs

Table 18 shows the complete joint-mode outputs for the case study in Table 7. The task asks to fill in the blank: "_____ is the direct attempt to formally or informally manage ethical issues or problems, through specific policies, practices and programmes." The assistant responds: "Let's think step by step. We refer to Wikipedia articles on business ethics for help. The direct attempt to manage ethical issues through specific policies, practices, and programs is business ethics management."

I Adapting SARA to Thinking Mode

Recent models support a *thinking mode* in which extended reasoning occurs inside a designated thinking block, and only the final answer appears in the visible output. This section describes how we adapt SARA's segment extraction and training to this setting, and reports preliminary results.

I.1 Differences from Instruct Mode

In instruct mode, SARA parses per-rubric segments from the structured output using a simple regex that matches the "`**Rubric` k `Analysis**: ... **Verdict**: ...`" pattern. Both the reasoning and the verdict are visible in the output, making segment boundaries unambiguous.

Thinking mode changes this in two ways:

1. **Reasoning moves to the thinking block.** The output prompt instructs the model to perform all

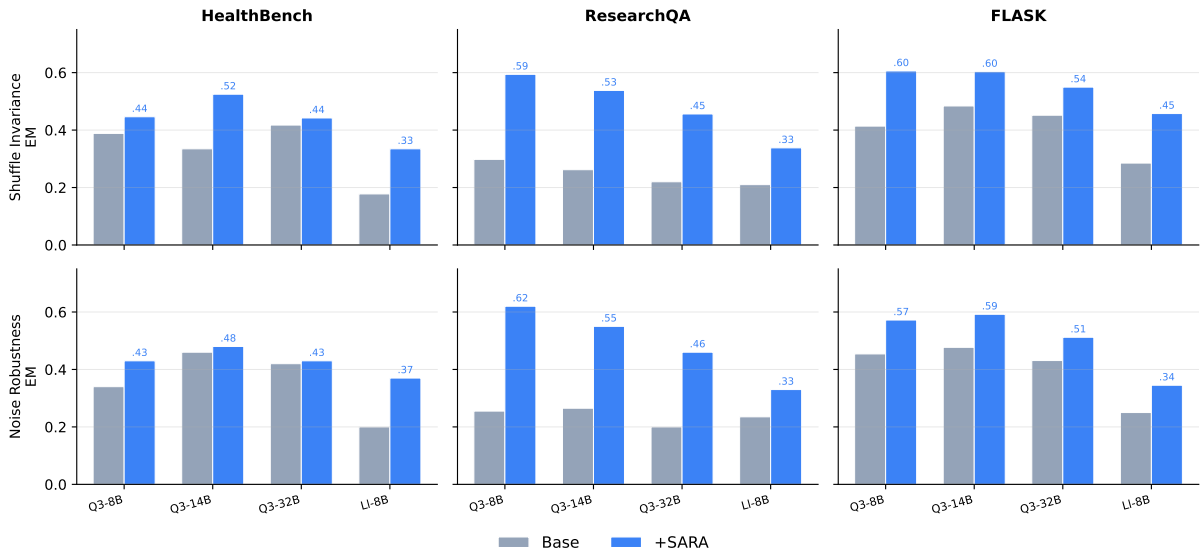

Figure 6: Sample-level exact match (EM) under shuffle invariance (top row) and noise robustness at 40% noise ratio (bottom row). Gray bars denote base models; blue bars denote SARA-trained models.

	HealthBench			ResearchQA			FLASK		
	Agr↑	κ↑	EM↑	Agr↑	κ↑	EM↑	Agr↑	κ↑	EM↑
Qwen3-8B	.886	.770	.388	.783	.695	.298	.709	.667	.414
+*SARA*	.911	.819	.446	.921	.879	.594	.839	.832	.606
Qwen3-14B	.888	.773	.335	.791	.746	.262	.732	.657	.484
+*SARA*	.934	.867	.525	.912	.886	.538	.831	.816	.603
Qwen3-32B	.900	.800	.417	.740	.733	.220	.703	.658	.452
+*SARA*	.908	.816	.442	.888	.868	.456	.862	.868	.651
Llama-3.1-8B	.783	.566	.178	.741	.626	.210	.604	.642	.285
+*SARA*	.887	.769	.335	.852	.780	.338	.772	.799	.458

Table 11: Shuffle invariance (5 permutations, averaged).

evaluation internally and produce only the final scores. The visible output reduces to a single line of scores (e.g., "`4+3+5+...=<sum>`").

2. **No structured segment boundaries.** The thinking block contains free-form reasoning without enforced formatting. Rubric analyses may be interleaved with reflection, planning, and self-correction steps.

I.2 Segment Extraction

Because the thinking block lacks rigid formatting, we use a paragraph-level heuristic. We split the thinking content into paragraphs (by double newlines) and scan each paragraph for rubric mentions using the pattern "`[Rr]ubric \s+(\d+)`". For each rubric k, we assign the *first* paragraph that mentions it as that rubric's analysis segment. Paragraphs that mention no specific rubric (typically planning, reflection, or summary paragraphs) are assigned to the structure mask.

This heuristic is less precise than instruct-mode parsing. A paragraph may discuss multiple rubrics, and some rubric reasoning may be distributed across non-contiguous paragraphs. We accept this noise as a trade-off for applicability.

I.3 Training Adjustments

We reduce the preservation loss weight from $\alpha = 0.4$ (instruct mode) to $\alpha = 0.2$. The thinking block contains substantial reflection and summary content beyond pure structural tokens. A high preservation weight on these paragraphs would overly constrain the model's reasoning process, reducing the distillation signal's ability to correct interference in the analysis segments.

I.4 Preliminary Results

We train SARA on Qwen3-14B in thinking mode using ResearchQA and compare against the instruct-mode results from the main experiments (Table 17).

Thinking-mode SARA improves over the base model (EM .23→.29, κ .658→.771) but substantially underperforms instruct-mode SARA (EM .52,

	HealthBench			ResearchQA			FLASK		
	Agr↑	κ↑	EM↑	Agr↑	κ↑	EM↑	Agr↑	κ↑	EM↑
Qwen3-8B	.876	.751	.340	.770	.663	.255	.740	.701	.454
+*SARA*	.908	.813	.430	.923	.877	.620	.826	.819	.572
Qwen3-14B	.896	.792	.460	.773	.712	.265	.737	.646	.477
+*SARA*	.924	.847	.480	.921	.880	.550	.819	.815	.592
Qwen3-32B	.892	.783	.420	.726	.714	.200	.694	.644	.431
+*SARA*	.905	.810	.430	.885	.861	.460	.847	.861	.615
Llama-3.1-8B	.755	.509	.200	.755	.621	.235	.589	.626	.250
+*SARA*	.873	.743	.370	.842	.772	.330	.713	.743	.345

Table 12: Noise robustness (40% noise ratio).

	ResearchQA			FLASK		
	MAD↓	RMSE↓	ρ↑	MAD↓	RMSE↓	ρ↑
Qwen3-8B	.273	.608	.816	.473	1.039	.728
+*SARA*	**.101**	**.361**	**.918**	**.289**	**.762**	**.862**
Qwen3-14B	.372	.844	.794	.441	.983	.743
+*SARA*	**.113**	**.385**	**.942**	**.247**	**.605**	**.889**
Qwen3-32B	.408	.832	.794	.471	.986	.737
+*SARA*	**.148**	**.472**	**.911**	**.243**	**.662**	**.822**
Llama-3.1-8B	.437	.919	.748	.542	.968	.763
+*SARA*	**.186**	**.535**	**.886**	**.364**	**.766**	**.854**

Table 13: Supplementary rubric-level metrics for overall consistency (scored datasets). MAD = mean absolute deviation, RMSE = root mean squared error, ρ = Pearson correlation.

κ .879). We attribute this gap to two factors:

Noisy segment boundaries. Paragraph-level extraction is coarser than regex-based parsing. Misaligned segments mean the distillation loss sometimes aligns the wrong tokens with isolation anchors, diluting the training signal.

Entangled reasoning. In thinking mode, the model often interleaves reasoning about multiple rubrics within a single paragraph, or revisits earlier rubrics during later reflection steps. This entanglement makes it harder to isolate per-rubric reasoning for targeted correction.

I.5 Discussion

These results suggest that SARA's core principle (using isolation judgments as anchors) applies across output formats, but the effectiveness depends heavily on segment extraction quality. Improving thinking-mode parsing, for example through training the model to structure its thinking block with explicit rubric delimiters, or through attention-based attribution methods, remains an open direction.

J Declaration of AI Assistant Usage

We explicitly disclose the nature and scope of AI assistance in this work:

- **Writing and Refinement:** We used Claude (Anthropic) strictly as a writing assistant to verify grammar, polish writing style, and refine the text of the manuscript.
- **Coding Assistance:** We employed Claude to draft routine boilerplate data processing scripts and infrastructure execution modules. However, the human authors entirely conceptualized and developed all core SARA algorithms, distillation losses, and custom pipeline code.
- **Human Ownership:** The human authors maintain sole ownership of all core scientific ideas, mathematical formulations, experimental designs, and data interpretations. We manually created all primary data visualizations, plots, and tables without any AI generation.

The use of GPT-4.1, as detailed in Section 5, serves strictly as an automated reference baseline/judge within our evaluation framework, rather than an interactive assistant in our research or writing process.

	ResearchQA			FLASK		
	TotalEM↑	TotalMAE↓	ρ_T↑	TotalEM↑	TotalMAE↓	ρ_T↑
Qwen3-8B	.250	1.620	.941	.431	1.316	.767
+*SARA*	**.640**	**0.470**	**.989**	**.557**	**0.695**	**.920**
Qwen3-14B	.250	2.100	.931	.437	1.206	.791
+*SARA*	**.610**	**0.550**	**.988**	**.609**	**0.592**	**.913**
Qwen3-32B	.310	2.340	.906	.425	1.264	.788
+*SARA*	**.530**	**0.720**	**.983**	**.615**	**0.626**	**.910**
Llama-3.1-8B	.210	2.100	.900	.305	1.316	.854
+*SARA*	**.390**	**0.950**	**.973**	**.386**	**0.454**	**.915**

Table 14: Aggregate-level metrics for overall consistency (scored datasets). TotalEM = fraction of samples with identical total scores, TotalMAE = mean absolute error of total scores, ρ_T = Pearson correlation of total scores.

Configuration	Agr↑	EM↑
EMA decay (β_d=0.5)		
decay = 1.0 (no EMA)	.890	.48
decay = 0.995	.911	.51
decay = 0.999 (default)	.900	.52
decay = 0.99	.918	.59
JSD interpolation β_d (decay=.999)		
β_d = 0.1	.905	.53
β_d = 0.3	.900	.54
β_d = 0.5 (default)	.900	.52
β_d = 0.7	.900	.52
β_d = 0.9	.867	.38

Table 15: Hyperparameter ablation (Qwen3-14B, ResearchQA, CONTRACTALL EM). Default: decay = .999, β_d=0.5.

	HB		RQA		FLASK	
	Base	+S	Base	+S	Base	+S
Qwen3-8B	405	611	453	691	230	290
Qwen3-14B	425	659	480	764	253	319
Qwen3-32B	466	762	501	821	262	594
Llama-3.1-8B	273	465	405	520	200	209
Avg Δ	+220		+259		+117	

Table 16: Average output tokens per sample (joint mode). +S = +SARA.

	Agr↑	κ↑	MAD↓	EM↑
Instruct mode				
Qwen3-14B	.768	.658	.372	.23
+SARA	.900	.879	.113	.52
Thinking mode				
Qwen3-14B	.765	.687	.323	.22
+SARA	.817	.771	.233	.29

Table 17: Overall consistency on ResearchQA. Thinking-mode SARA improves over the base but underperforms instruct-mode SARA.

Base Model (joint)	+SARA (joint)
R1 Analysis: The response attempts to address the instruction by identifying the concept as "business ethics management." However, it does not fully answer the question, which is asking for the term that fills in the blank. The response is somewhat relevant but misses the key requirement of directly providing the term that completes the sentence. **Verdict: 4/5.** (iso: 3)	**R1 Analysis**: The response attempts to address the question by identifying "business ethics management" as the concept being described. However, it does not fully or clearly define the term, nor does it provide a complete or precise answer to the question. The response is somewhat relevant but lacks the clarity and completeness expected. **Verdict: 3/5.** (iso: 3)
R2 Analysis: The response does not provide any factual evidence or citation to support the claim that the term is "business ethics management." While the concept is plausible, there is no reliable source or background knowledge explicitly referenced to confirm its accuracy. **Verdict: 2/5.** (iso: 3)	**R2 Analysis**: The assistant's response states that the direct attempt to manage ethical issues through specific policies, practices, and programs is "business ethics management." However, the response lacks any citation or reference to a reliable source such as Wikipedia or any other authoritative text. While the concept may be accurate, the support for the response is incomplete and not fully reliable. **Verdict: 3/5.** (iso: 3)
R3 Analysis: The response is short and straightforward, with minimal ambiguity. It is generally understandable, though it could be more clearly structured to directly answer the question by providing the term in the blank. **Verdict: 4/5.** (iso: 4)	**R3 Analysis**: The response is clear and well-structured, beginning with a statement that acknowledges the need to think step by step and referencing a source (Wikipedia articles on business ethics) for support. The answer directly addresses the question and provides a concise explanation. The organization is logical and easy to follow. **Verdict: 5/5.** (iso: 5)
Total: 4+2+4=10	**Total: 3+3+5=11**

Table 18: Full joint-mode outputs for the case study. The highlighted region in the base model's R2 shows absolutist phrasing ("does not provide *any*") absent from its own isolation output, where it acknowledges partial accuracy. Parenthetical annotations show each model's isolation verdict for comparison. SARA matches isolation on all three rubrics; the base model deviates on R1 (+1) and R2 (–1).