WHEN RICHER PREFERENCES DO NOT HELP: RUBRIC REDUNDANCY, PAIR SELECTION, AND TRAINING EXPOSURE IN SCIENTIFIC QA

Anonymous authors
Paper under double-blind review

ABSTRACT

Fine-grained rubrics and dense preference pairs appear to offer richer supervision for aligning scientific-domain large language models (LLMs), but nominal supervision richness need not translate into non-redundant usable preference information. We study this question on 4,000 titanium-alloy reverse-design questions, each associated with four scored candidate answers and task-specific rubric vectors. In the audited single-checkpoint matrix, clean-start binary and six-pair graded DPO both display approximately 92.6 mean generated-answer score at one-decimal precision. Rubric dimensions are strongly correlated for several reasoning types, and at q50 rubric-absolute and scalar-gap pair selection overlap substantially (Jaccard 0.943), although the scalar score is itself aggregated from the rubric dimensions. Across three training seeds, rubric-aware q50 filtering does not establish a reliable advantage over scalar or random selection. Moreover, a seed-42 truncated full-data probe reaches 92.85 after approximately 9.7k preference-pair exposures, compared with 92.31 after a full pass over 19.2k pairs, showing that one-epoch filtered-versus-full comparisons change training exposure together with pair composition. Together, these results suggest that, in this controlled setting, nominal supervision richness can be a poor proxy for non-redundant preference information; preference-signal efficiency should therefore be analyzed through feedback redundancy, pair composition, and training exposure.

1 INTRODUCTION

Preference optimization is widely used to align large language models (LLMs) with human or AI feedback. A natural assumption is that richer feedback should improve alignment: a graded answer ladder supplies more ordered pairs than a binary comparison, and rubric vectors expose more detail than a scalar preference. This assumption is especially appealing in scientific domains, where answer quality is inherently multidimensional. A response may be factually plausible but mechanistically shallow, technically sound but poorly supported, or diagnostically correct but operationally incomplete.

Scientific and materials benchmarks nevertheless show that domain-specific reasoning, evidence use, and specialized knowledge remain challenging for general-purpose LLMs (Wang et al., 2023; Song et al., 2023; Zaki et al., 2023; Zhang et al., 2025). In titanium-alloy reverse design, models must reason across composition–processing–microstructure–property relations, answer causal and counterfactual questions, explain mechanism pathways, and diagnose engineering failures. This setting provides a useful test bed for asking whether additional preference structure contributes genuinely new alignment information or mostly repeats the same underlying quality ordering.

We study a rubric-scored QA resource in which every question has four candidate answers and both scalar and dimension-level scores. This structure supports binary and six-pair graded preference optimization from the same answer groups, while the dimension-level scores enable rubric-aware analyses of preference-pair selection. In the audited single-checkpoint matrix, denser pair supervision does not yield a consistent gain across the examined starting checkpoints. Rather than treating this observation as a binary-versus-graded verdict, we use it to ask a broader question: when does

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

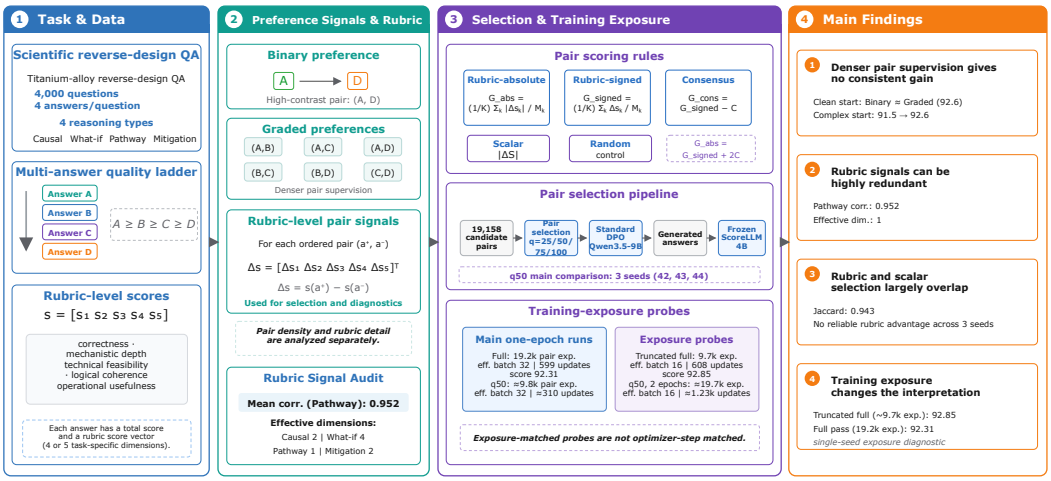

Figure 1: Overview of preference signal efficiency in scientific-domain QA alignment.

nominally richer supervision contribute non-redundant usable preference information rather than repeat an existing quality ordering?

The same answer ladder also exposes a training-exposure confound that is easy to miss when supervision formats are compared by nominal epochs. A binary condition can contribute one pair for a question, whereas a graded condition contributes six; training both for one epoch therefore changes not only which preference relations are available but also the amount of training exposure and, depending on the batch configuration, the number of optimizer updates. Likewise, filtering half of a dense pair pool changes both pair composition and the number of preference-pair examples processed. For this reason, we treat feedback richness, pair selection, and training exposure as separate experimental factors rather than interchangeable notions of "more" or "less" supervision.

The central question of this paper is what the richer rubric information is actually doing. We organize the study around three research questions. **RQ1:** Does moving from high-contrast binary to denser six-pair graded supervision yield a consistent alignment gain under a common evaluation? **RQ2:** How much pair-selection information do rubric dimensions provide beyond a scalar score gap? **RQ3:** When filtered subsets appear competitive with full preference data, how much of the effect can be separated from reduced training exposure?

Our experiments connect these questions, as summarized in Figure 1. First, a rubric audit characterizes correlation and effective dimensionality across reasoning types. We then use the same rubric vectors as pair-selection signals, comparing absolute, signed, and conflict-penalized rubric gaps with scalar-gap and random controls. Finally, we add multi-seed q50 comparisons and approximate training-exposure probes. These probes show that filtered-data results cannot be interpreted from pair identity alone: a truncated full-data run at a comparable pair-exposure level is itself a competitive comparator under the evaluator.

Contributions. Our study makes three empirical contributions: **(1)** an audited comparison of high-contrast binary and six-pair graded preference supervision, showing that denser pair supervision does not provide a consistent gain across the examined settings; **(2)** a rubric-information audit combining dimension correlation, effective dimensionality, selected-set overlap, and direction conflicts to assess how much pair-selection information rubric vectors add beyond scalar gaps; and **(3)** multi-seed selection experiments and training-exposure probes showing that apparent preference-data efficiency must be interpreted jointly with pair composition and training exposure. Together, these analyses study when nominally richer supervision contributes distinct usable information rather than assuming that more annotations or pairs are inherently better.

2 Related Work

Scientific-domain LLMs and materials reasoning. Scientific-domain LLMs require reliable reasoning over specialized concepts and constraints. Existing scientific and materials benchmarks evaluate scientific problem solving, materials-language understanding, and materials QA (Wang et al., 2023; Song et al., 2023; Zaki et al., 2023; Zhang et al., 2025). They motivate domain-specific evaluation, but mainly target answer correctness or benchmark accuracy. We instead study how preference-supervision structure affects scientific-domain alignment.

Preference optimization and data-centric alignment. LLM alignment commonly uses SFT and RLHF (Ouyang et al., 2022), while DPO directly optimizes chosen–rejected pairs without a separate reward model (Rafailov et al., 2023). Later work explores rank-based, reference-free, guiding-reference, and high-confidence objectives (Yuan et al., 2023; Zhao et al., 2023; Ethayarajh et al., 2024; Meng et al., 2024; Hong et al., 2024; Pan et al., 2026; Huang et al., 2026), while data-centric studies emphasize data quality, selection, noise, and preference information content (Zhou et al., 2023; Cui et al., 2023; Ivison et al., 2024; Chen et al., 2024; Shen et al., 2024; Deng et al., 2026; Pan et al., 2025). Multi-level preference learning has exploited adjacent and cross-level comparisons in MLLMs (Zhang et al., 2024), and preference-difference filtering shows that larger, more consistent gaps can improve reward modeling and DPO (Lin et al., 2025). These works motivate stronger preference construction, but do not isolate whether richer rubric supervision adds incremental information once pair composition and training exposure are separated. We study this distinction in scientific QA.

Rubric-based evaluation and multi-objective alignment. Open-ended generation motivates LLM-as-a-judge and rubric-based evaluation (Zheng et al., 2023; Liu et al., 2023; Hashemi et al., 2024; Fan et al., 2024; Lambert et al., 2025). Multi-attribute and multi-objective alignment represents diverse dimensions rather than collapsing feedback into one scalar reward (Wang et al., 2024a;b;c; Zhou et al., 2024; Yang et al., 2024; Gupta et al., 2025; Lou et al., 2025). These studies establish the value of decomposed feedback, but leave open when rubric dimensions add information beyond aggregate scores. Our rubric audit addresses this through dimension correlation, effective dimensionality, and rubric-versus-scalar pair-selection overlap.

3 Preference Signals and Pair Selection

3.1 Answer Ladder and Feedback Richness

We use a titanium-alloy reverse-design QA resource with four reasoning types: *causal*, *what-if*, *pathway*, and *mitigation*. Each question q_i has four candidate answers $\{a_i^A, a_i^B, a_i^C, a_i^D\}$ obtained by stable sorting on the dataset's original total score,

$$S(a_i^A) \geq S(a_i^B) \geq S(a_i^C) \geq S(a_i^D). \tag{1}$$

The weak inequality matters because 183 training questions yield 184 zero-gap pair positions (0.96% of the 19,158 graded pairs); ties inherit the stable input order. The six ranked positions are (A, B), (A, C), (A, D), (B, C), (B, D), and (C, D).

The *binary* condition retains only top–bottom pairs with $\Delta S \geq 15$, yielding 2,529 pairs and 79.2% training-question coverage, versus 19,158 pair positions for the graded condition. Thus, the comparison changes pair density, margin distribution, question coverage, role frequency, and training exposure, and is treated as a supervision-format rather than an isolated pair-density intervention. In the graded construction, A and D appear three times exclusively as chosen and rejected, respectively, while B and C appear on both sides. Appendix D examines pair-position sensitivity.

3.2 Rubric-Aware Selection Scores

We next ask whether rubric vectors are more useful for deciding *which* preference pairs to train on. For rubric dimension k with maximum score M_k, define the normalized signed gap

$$d_k = \frac{s_k^+ - s_k^-}{M_k}. \tag{2}$$

We use three rubric-derived selection scores:

$$G_{\text{signed}} = \frac{1}{K} \sum_k d_k, \tag{3}$$

$$C = \frac{1}{K} \sum_k \max(0, -d_k), \tag{4}$$

$$G_{\text{abs}} = \frac{1}{K} \sum_k |d_k|, \tag{5}$$

$$G_{\text{cons}} = G_{\text{signed}} - C. \tag{6}$$

G_{abs} rewards a large dimension-wise separation regardless of direction, whereas G_{signed} preserves the chosen–rejected direction. C measures dimensions that favor the rejected answer, and G_{cons} penalizes such conflicts. We compare these rules with the scalar total-score gap $G_{\text{scalar}} = |\Delta S|$ and random selection.

For a score G and a nominal retention budget $q \in \{25, 50, 75, 100\}$, we retain approximately the top $q\%$ of pairs according to G and then enforce at least one retained pair per question. We denote these settings as q25, q50, q75, and q100. Because the per-question coverage constraint can add pairs beyond the nominal threshold, q is a target rather than an exact retained fraction: at q50, scalar selection keeps 9,923 of 19,158 training pairs (51.8%), while consensus keeps 9,838 (51.4%).

The control rules answer different questions. Scalar-gap selection tests whether dimension-level information adds value beyond the aggregate score from which the answer ladder is derived; random selection tests whether simply reducing the pair pool is sufficient; and signed/consensus variants isolate the effect of rubric dimensions that disagree with the scalar chosen–rejected direction. Because the per-question coverage constraint is applied to every filtered condition, q50 reduces pair density while preserving representation of all 3,193 training questions.

3.3 Algebraic and Collinearity Properties

The rubric-gap family obeys the following bookkeeping identity:

$$\boxed{G_{\text{abs}} = G_{\text{signed}} + 2C.} \tag{7}$$

It follows directly from $|x| = x + 2\max(0, -x)$ in each normalized dimension. Although algebraically elementary, the identity makes explicit how the selection rules treat direction conflicts: absolute-gap selection assigns them a positive contribution, whereas consensus selection penalizes them. This relation does not depend on how the scalar total score is constructed.

A complementary limiting case explains why rubric and scalar selection can become redundant. For a fixed task type, suppose every rubric dimension is a positive affine function of one latent quality variable, $s_k(a) = \alpha_k L(a) + b_k$ with $\alpha_k > 0$, and the total score is a positive linear combination of dimensions. Then every strictly ordered pair has $C = 0$, and the rubric-gap scores and scalar total gap are positive multiples of the same latent difference, inducing the same ranking up to ties. Our data are not perfectly collinear, and mixed task types have different normalizations; therefore this is a limiting explanation rather than a quantitative prediction of the observed selected-set overlap.

4 Experimental Setup

4.1 Dataset Construction and Splits

The preference experiments are built on a provenance-aware titanium-alloy QA-score resource rather than on isolated prompt–response pairs. Heterogeneous literature-derived material records are first governed into a retrieval-ready corpus by normalizing units and terminology while preserving source identifiers. Type-specific structured scenarios are then instantiated for causal, what-if, pathway, and mitigation reasoning. Retrieval is used as construction-time grounding: records seed scenario fields and supply compact evidence for answer generation, rather than serving as an independently evaluated open-domain retrieval benchmark. Multiple answer sources provide four

Table 1: Question-level split by reasoning type.

Type	Dim.	Train	Val.	Test
Causal	4	797	98	105
What-if	5	798	98	104
Pathway	4	799	99	102
Mitigation	4	799	99	102
Total	–	3193	394	413

candidate responses per question, and a two-stage task-specific rubric procedure produces both total scores and dimension-level score profiles.

The construction keeps a lightweight provenance chain linking source/evidence identifiers, the structured scenario, question, answer-source labels, and score records. This provenance is useful for the present study because binary, graded, and rubric-aware pair-selection variants can all be derived from the same question groups without changing the underlying scientific scenario or test-question distribution. Appendix A gives the construction stages, instance schema, scenario fields, and rubric definitions in more detail.

The resource contains 56,515 literature-derived retrieval records, 4,000 questions, 16,000 candidate answers, and 32,000 two-stage scoring records. Splits are made at the question level, retaining all four answers of a question within the same split. Table 1 reports the resulting 3,193/394/413 train/validation/test questions. The training split yields $3,193 \times \binom{4}{2} = 19,158$ ranked pair positions before filtering.

4.2 Audited Common Evaluation

All generator checkpoints are evaluated on the same 413 test questions. Generation uses temperature 0.3, top-$p = 0.95$, and at most 1,024 new tokens; one sampled generation per question is cached and reused for all reported metrics. A frozen Qwen3.5-4B-based ScoreLLM greedily emits task-specific rubric scores and a total score. The primary metric is the mean generated-answer score

$$\bar{S}_{\text{gen}} = \frac{1}{413} \sum_{i=1}^{413} S_{\text{ScoreLLM}}(q_i, \hat{a}_i). \tag{8}$$

This metric keeps all compared generations on the same automatic scoring scale. Metrics relative to dataset-original A/B/C/D labels are relegated to the appendix because those labels and generated-answer ScoreLLM scores are not guaranteed to share an identical numerical scale. On held-out scoring annotations, the frozen evaluator reports MAE/RMSE of 0.910/1.673 and Pearson/Spearman correlations of 0.9649/0.9625. We use these figures only as evidence that the evaluator reproduces its domain scoring annotations reasonably well; they are not a substitute for independent human evaluation of the newly generated answers.

We align evaluation outputs by unique question ID, verify the expected per-type counts (105/104/102/102), and explicitly map generation caches to score records. All main-text results use the same robustly parsed evaluation pipeline. Parser failure modes, cache reconciliation, and audit details are reported in Appendix B.

4.3 Selection and Training-Exposure Probes

One-epoch filtering changes both pair composition and total training exposure. The main seed-42 full, scalar-q50, and consensus-q50 runs use effective batch size 32 and process 19,158, 9,923, and 9,838 preference pairs, corresponding to 599, 311, and 308 optimizer updates, respectively.

To probe whether the one-epoch comparison is confounded by exposure, we add three seed-42 runs using effective batch size 16. A truncated full-data run processes approximately 9,728 pair examples in 608 optimizer updates; scalar q50 is trained for two epochs, producing 19,846 pair exposures in 1,242 updates; and consensus q50 is trained for two epochs, producing 19,676 pair exposures in 1,230 updates. We count repeated pairs again when they are revisited in later epochs.

Table 2: Audited preference-richness matrix. Values are mean generated-answer ScoreLLM scores and matrix conditions are single checkpoints. IDs retain the original run-registry labels; the validated full matrix and intermediate iterative checkpoints are reported in Appendix C.

ID	Configuration	$\bar{S}_{\text{gen}} \uparrow$
M0	Zero-shot	75.6
M1	Clean SFT	86.8
M2	Clean-start binary DPO	92.6
M3	Clean-start graded DPO	92.6
M4	Complex-start binary DPO	91.5
M5	Complex-start graded DPO	92.6
M10	Iterative final	92.4

These are approximate exposure controls rather than step-matched interventions: the main and probe runs use different effective batch sizes, the truncated full-data run observes only an early shuffled portion of the full pool, and the two-epoch filtered runs repeat retained pairs.

4.4 Training Configuration

All generator models use Qwen3.5-9B. Plain SFT uses LoRA with rank 8 and scale 16 and is trained for three epochs on high-quality answers. Standard DPO uses 4-bit NF4 QLoRA, bf16 computation, $\beta = 0.1$, learning rate 5×10^{-5}, cosine scheduling, 3% warmup, and maximum sequence length 768. Batch configuration varies across experiment families: the main one-epoch pair-selection runs use effective batch size 32, whereas the training-exposure probes use effective batch size 16. Run-specific exposure and optimizer-update counts are summarized in Section 4.3 and analyzed in Section 5.4.

We use *clean-start* for DPO initialized from Plain SFT and *complex-start* for DPO initialized from the Complex SFT checkpoint described in Appendix B. OrthoAux denotes the corresponding orthogonality-focused auxiliary checkpoint, while "iterative final" refers to the final checkpoint of the iterative sequence whose intermediate checkpoints are reported in Appendix C. Full objectives and hyperparameters are provided in Appendix B.

5 Results

5.1 Denser Pair Supervision Does Not Yield a Consistent Gain

Table 2 reports the central single-checkpoint conditions from the audited preference-richness matrix. Under the clean start, binary and six-pair graded DPO both display 92.6 at one-decimal precision. Because these are single trained checkpoints, this descriptive equality should not be interpreted as evidence of statistical equivalence. Under the complex start, binary and graded DPO obtain 91.5 and 92.6, respectively, while the iterative final checkpoint remains near 92.4.

Thus, graded DPO matches binary DPO at the reported precision under the clean start and exceeds it under the complex start, but neither comparison is replicated across training seeds. The current matrix therefore does not establish a consistent performance gain from denser pair supervision.

The aggregate scores also mask descriptive task-level variation. For example, clean-start binary versus graded DPO scores 87.3 versus 86.6 on what-if questions, while graded DPO scores 93.5 versus 92.2 for binary DPO on pathway questions. These are single-checkpoint means rather than replicated type-specific effects, so they do not justify an adaptive routing rule. The corresponding per-type table is reported in Appendix C.

5.2 Rubric Redundancy and Limited Incremental Selection Information

Table 3 audits answer-level rubric structure using only the training split. Within each reasoning type, each scored answer is one observation; we compute the mean pairwise Pearson correlation across rubric dimensions and apply PCA after dimension-wise standardization. Effective dimensionality

Table 3: Training-split answer-level rubric audit. Mean r is the mean pairwise Pearson correlation; PCA uses standardized dimensions. Eff. Dim. is the minimum number of components explaining at least 95% of the variance.

Type	K	Mean r	PC1 Var.	Eff. Dim.
Causal	4	.834	87.7%	2
What-if	5	.580	67.4%	4
Pathway	4	.952	96.4%	1
Mitigation	4	.899	92.4%	2

Table 4: Nominal q50 selection and full-data training across three seeds. Sample SD ($\mathrm{ddof} = 1$) is descriptive over the three runs.

Method	42	43	44	Mean $\pm$ SD
Full six pairs	92.31	92.92	91.75	92.33$\pm$.59
Random q50	92.68	92.35	92.76	92.60$\pm$.22
Scalar q50	93.04	91.61	92.41	92.35$\pm$.72
Rubric-abs q50	92.82	93.04	92.48	92.78$\pm$.28
Consensus q50	92.91	92.84	92.70	92.82$\pm$.11

is the minimum number of principal components required to explain at least 95% of cumulative variance. Pathway, mitigation, and causal scores are strongly correlated, whereas what-if scores are less redundant. For pathway, PC1 explains 96.4% of training-score variance and the effective dimensionality is one.

The pair-selection data show similarly limited divergence. At q50, scalar-gap and rubric-absolute selection have Jaccard overlap 0.943. Because the scalar total is itself aggregated from the rubric dimensions, substantial overlap is expected; the Jaccard value therefore quantifies selected-set similarity rather than independently proving rubric redundancy. A pair-gap sensitivity analysis lowers absolute correlations but preserves the qualitative pattern: pathway remains highly correlated and what-if the least correlated type (Appendix D.3).

The rubric variants nevertheless change one interpretable property: selected-pair conflict rates are 3.9% for absolute gap, 1.9% for signed gap, and 1.4% for consensus, in agreement with the algebraic role of C in Eq. 7. These composition differences show that rubric-aware rules select some genuinely different examples; they do not by themselves establish a performance benefit.

5.3 Three Seeds Do Not Establish a Rubric-Selection Advantage

Table 4 compares the principal q50 rules across seeds 42–44. Consensus has the highest observed mean (92.82) and the smallest observed sample standard deviation (0.11), but three seeds are too few to infer a general variance ordering. Rubric-absolute reaches 92.78, random q50 reaches 92.60, scalar q50 reaches 92.35, and full training reaches 92.33. The matched-seed consensus-minus-full differences are $+0.60$, -0.08, and $+0.95$, including one negative result. We therefore do not infer that rubric-aware selection is reliably superior to scalar, random, or full-data training.

The single best seed is also misleading: scalar q50 is highest at seed 42 but has a lower three-seed mean than both rubric variants and random q50. Because training randomness, sampled generation, and random-selection randomness are coupled to run seeds, these experiments support a limited claim: several distinct q50 subsets can remain competitive under the evaluator, while a distinct rubric advantage has not been established. Importantly, the coverage constraint means that these q50 conditions still contain at least one pair from every training question. Their differences therefore arise from pair density and pair composition within a fixed question set rather than from dropping half of the scientific questions altogether.

5.4 Training Exposure Changes the Interpretation of Filtering Gains

The seed-42 probes in Table 5 expose a central training-exposure confound. Scalar q50 processes 9,923 pair examples and reaches 93.04, whereas the truncated full-data probe processes approxi-

Table 5: Seed-42 training-exposure probes. Pair exposures count processed preference-pair examples, including repeats across epochs. Main one-epoch runs use effective batch size 32, whereas the exposure probes use 16; the comparisons are therefore not optimizer-step matched.

Data / duration	Pair exp.	Eff. batch	Steps	$\bar{S}_{\text{gen}}$
Full, one epoch	19,158	32	599	92.31
Full, truncated	9,728	16	608	92.85
Scalar q50, one epoch	9,923	32	311	93.04
Scalar q50, two epochs	19,846	16	1,242	91.71
Consensus q50, one epoch	9,838	32	308	92.91
Consensus q50, two epochs	19,676	16	1,230	92.82

mately 9,728 pair examples and reaches 92.85. In contrast, full one-epoch training processes 19,158 pair examples and reaches 92.31. Thus, the original one-epoch q50-versus-full comparison changes training exposure together with pair identity and does not isolate a pair-selection effect.

At roughly matched total pair exposure ($\sim$19–20k), full one-epoch training obtains 92.31, scalar q50 trained for two epochs obtains 91.71, and consensus q50 trained for two epochs obtains 92.82. These runs are exposure-matched only approximately: the filtered two-epoch runs repeat the same retained pairs and use a smaller effective batch, resulting in roughly twice as many optimizer updates as the full one-epoch run. Consequently, the differences cannot be uniquely attributed to pair selection, repeated exposure, or update count. The defensible conclusion is that training exposure changes the interpretation of pair-filtering gains in this single-seed diagnostic and should be treated as part of the experimental condition.

6 Discussion

From nominal richness to usable information. The experiments separate three factors that are easily conflated: adding pairs changes training constraints, rubric decomposition changes annotation detail, and filtering changes both selected examples and training exposure. In this controlled setting, nominal supervision richness can therefore be a poor proxy for non-redundant usable preference information. The lack of a consistent gain from denser pairing, redundant answer-level rubric structure, and strong scalar/rubric selected-set overlap are jointly consistent with limited incremental information, while the pair-gap sensitivity analysis shows that the strength of rubric correlation depends on the analysis unit. The multi-seed q50 results nevertheless do not establish a robust ordering among selection rules.

Rubrics remain useful without an established selection advantage. Rubric vectors expose dimension conflicts hidden by scalar scores, and Eq. 7 shows how absolute, signed, and consensus rules treat them differently. This makes rubrics useful for diagnosis and controlled selection even though the current multi-seed experiments do not establish a reproducible advantage over scalar selection. The secondary SADD study likewise finds no reliable evidence that more elaborate rubric weighting improves over simpler weighting schemes (Appendix F).

Pair position is not the same as score contrast. In the exploratory seed-42 leave-one-pair analysis (Appendix D), removing (A, B) causes the largest degradation (90.01), whereas removing (A, D) and (B, C) yields 92.07 and 92.72. This challenges a simple largest-gap explanation, but because removal also changes chosen/rejected role frequencies, we treat the result only as hypothesis-generating evidence rather than a ranking of intrinsic pair utility.

Implications for preference-data efficiency. One-epoch comparisons confound pair composition with training exposure. Preference-data efficiency studies should therefore report retained-pair counts, total pair-example exposures, effective batch size, and optimizer updates rather than epochs alone. When possible, exposure-matched, step-matched, or learning-curve controls should be included explicitly. Our current probes show that one-epoch scores do not isolate selection quality, but they do not factorially separate pair identity, repeated exposure, batch size, and optimizer-update count.

Implications for preference-dataset design. Multi-answer datasets should expose enough structure to audit how supervision is constructed. Answer ladders distinguish adjacent from high-contrast comparisons, while rubric vectors expose disagreement among quality dimensions. Before adding dimensions or pairs, dataset builders can inspect correlation, selected-set overlap, score margins, and answer-role frequencies to determine whether richer annotations materially change the training signal.

7 LIMITATIONS

The study is limited to one scientific domain, one main generator backbone, and a domain-adapted ScoreLLM linked to the preference-scoring pipeline. The evaluator may favor answers consistent with its training annotations, so the results do not establish independent human-preference improvements or cross-domain validity. No independent human evaluation was conducted on the revised generated-model outputs; the sampled human consistency checks in our data pipeline apply to dataset construction rather than to the model comparisons reported here. Because splitting is performed at the question level, residual source- or evidence-level overlap across splits cannot be completely ruled out.

The main q50 selection comparison has only three seeds, while the training-exposure probes, leave-one-pair analysis, SADD variants, and most q25/q75 conditions have one seed. Fixed-checkpoint paired bootstrap results in the appendix measure variation over the 413 matched questions but do not incorporate training-seed uncertainty and are not used as method-level significance claims.

Several analyses are exploratory. The q50 emphasis, the fixed consensus selection rule, the long-training consensus interpretation, and the pair-position diagnostics were examined on the same test set used for reporting. They should be validated with prespecified settings on an independent holdout or external dataset. In addition, dataset-original reference scores and ScoreLLM scores for newly generated answers arise from different scoring protocols; cross-scale reference-relative metrics are therefore secondary rather than central evidence.

The training-exposure probes are approximate exposure controls rather than optimizer-step-matched interventions. The main one-epoch runs use effective batch size 32, whereas the truncated and two-epoch probes use effective batch size 16. Thus, the truncated full-data run approximately matches the q50 conditions in processed pair examples ($\sim$9.7–9.9k) but not in optimizer updates, while the two-epoch filtered runs approximately match full-data exposure ($\sim$19–20k) but repeat the same retained pairs and use substantially more updates. Learning-rate schedules and warmup also differ with run length. More complete factorial controls over pair identity, distinct-example count, repeated exposure, effective batch size, and optimizer updates are needed to isolate these effects.

Finally, this work should be viewed as a controlled methodological study at the experimental stage rather than an industrial alloy-design or deployment evaluation. The generated answers are research artifacts used to study preference alignment behavior; they have not been experimentally validated as alloy-design recommendations and should not be used as direct guidance for materials design or engineering deployment.

8 CONCLUSION

The current audited matrix does not establish a consistent gain from denser preference-pair supervision in this scientific QA setting. Rubric redundancy is substantial but analysis-unit dependent, rubric-absolute and scalar-gap selection largely overlap, and three-seed q50 experiments do not establish a reliable rubric-selection advantage. At the same time, a single-seed truncated full-data probe at approximately matched pair exposure changes the interpretation of filtered-data gains by showing that pair composition and training exposure vary together. Taken together, these findings suggest that, in this controlled setting, nominal supervision richness can be a poor proxy for non-redundant usable preference information. Preference-signal efficiency should therefore be analyzed jointly through feedback redundancy, pair composition, and training exposure rather than annotation dimensionality, pair count, or retained-data fraction alone.

AI USE STATEMENT

Generative AI tools were used in the research workflow for the LLM-assisted construction of scientific question–answer instances, multi-model answer generation, and rubric-based automatic scoring described in the paper and appendix. Generative AI was also used for language and formatting assistance during manuscript preparation. The authors reviewed the resulting data, analyses, claims, and manuscript, performed the reported validation and statistical checks, and take responsibility for the final content.

REPRODUCIBILITY STATEMENT

The main paper and appendix document the dataset schema and question-level splits, task-specific rubrics, preference-pair construction, pair counts, optimization settings, random seeds, evaluation protocol and audit, pair-selection rules, and training-exposure probes. Additional implementation and diagnostic details are provided in the appendix. An anonymized code package containing the releasable training, evaluation, analysis, and configuration files is submitted as supplementary material. The full structured dataset remains subject to project-specific release constraints.

REFERENCES

Angelica Chen, Sadhika Malladi, Lily H Zhang, Xinyi Chen, Qiuyi Zhang, Rajesh Ranganath, and Kyunghyun Cho. Preference learning algorithms do not learn preference rankings. *Advances in Neural Information Processing Systems*, 37:101928–101968, 2024.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, et al. Ultrafeedback: Boosting language models with scaled ai feedback. *arXiv preprint arXiv:2310.01377*, 2023.

Xun Deng, Han Zhong, Rui Ai, Fuli Feng, Zheng Wang, and Xiangnan He. Less is more: Improving llm alignment via preference data selection. *Advances in Neural Information Processing Systems*, 38:161259–161285, 2026.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. Model alignment as prospect theoretic optimization. In *Forty-first International Conference on Machine Learning*, 2024.

Zhiyuan Fan, Weinong Wang, Debing Zhang, et al. Sedareval: Automated evaluation using self-adaptive rubrics. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 16916–16930, 2024.

Raghav Gupta, Ryan Sullivan, Yunxuan Li, Samrat Phatale, and Abhinav Rastogi. Robust multi-objective preference alignment with online dpo. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 27321–27329, 2025.

Helia Hashemi, Jason Eisner, Corby Rosset, Benjamin Van Durme, and Chris Kedzie. Llm-rubric: A multidimensional, calibrated approach to automated evaluation of natural language texts. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13806–13834, 2024.

Jiwoo Hong, Noah Lee, and James Thorne. Orpo: Monolithic preference optimization without reference model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 11170–11189, 2024.

Yue Huang, Xiangqi Wang, and Xiangliang Zhang. Spa: Achieving consensus in llm alignment via self-priority optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 31220–31228, 2026.

Hamish Ivison, Yizhong Wang, Jiacheng Liu, Zeqiu Wu, Valentina Pyatkin, Nathan Lambert, Noah A Smith, Yejin Choi, and Hannaneh Hajishirzi. Unpacking dpo and ppo: Disentangling best practices for learning from preference feedback. *Advances in neural information processing systems*, 37:36602–36633, 2024.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, et al. Rewardbench: Evaluating reward models for language modeling. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 1755–1797, 2025.

Qi Lin, Hengtong Lu, Caixia Yuan, Xiaojie Wang, Huixing Jiang, and Wei Chen. Data with high and consistent preference difference are better for reward model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 27482–27490, 2025.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 2511–2522, 2023.

Xingzhou Lou, Junge Zhang, Jian Xie, Lifeng Liu, Dong Yan, and Kaiqi Huang. Sequential preference optimization: Multi-dimensional preference alignment with implicit reward modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 27509–27517, 2025.

Yu Meng, Mengzhou Xia, and Danqi Chen. Simpo: Simple preference optimization with a reference-free reward. *Advances in Neural Information Processing Systems*, 37:124198–124235, 2024.

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

Junshu Pan, Wei Shen, Shulin Huang, Qiji Zhou, and Yue Zhang. Pre-dpo: Improving data utilization in direct preference optimization using a guiding reference model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 32646–32654, 2026.

Yu Pan, Zhongze Cai, Huaiyang Zhong, Guanting Chen, and Chonghuan Wang. What matters in data for dpo? *Advances in Neural Information Processing Systems*, 38:49957–49984, 2025.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Judy Hanwen Shen, Archit Sharma, and Jun Qin. Towards data-centric rlhf: Simple metrics for preference dataset comparison. *arXiv preprint arXiv:2409.09603*, 2024.

Yu Song, Santiago Miret, and Bang Liu. Matsci-nlp: Evaluating scientific language models on materials science language tasks using text-to-schema modeling. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3621–3639, 2023.

Haoxiang Wang, Yong Lin, Wei Xiong, Rui Yang, Shizhe Diao, Shuang Qiu, Han Zhao, and Tong Zhang. Arithmetic control of llms for diverse user preferences: Directional preference alignment with multi-objective rewards. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8642–8655, 2024a.

Haoxiang Wang, Wei Xiong, Tengyang Xie, Han Zhao, and Tong Zhang. Interpretable preferences via multi-objective reward modeling and mixture-of-experts. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 10582–10592, 2024b.

Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. Scibench: Evaluating college-level scientific problem-solving abilities of large language models. *arXiv preprint arXiv:2307.10635*, 2023.

Zhilin Wang, Yi Dong, Olivier Delalleau, Jiaqi Zeng, Gerald Shen, Daniel Egert, Jimmy J Zhang, Makesh N Sreedhar, and Oleksii Kuchaiev. Helpsteer 2: Open-source dataset for training top-performing reward models. *Advances in Neural Information Processing Systems*, 37:1474–1501, 2024c.

Rui Yang, Xiaoman Pan, Feng Luo, Shuang Qiu, Han Zhong, Dong Yu, and Jianshu Chen. Rewards-in-context: Multi-objective alignment of foundation models with dynamic preference adjustment. *arXiv preprint arXiv:2402.10207*, 2024.

Zheng Yuan, Hongyi Yuan, Chuanqi Tan, Wei Wang, Songfang Huang, and Fei Huang. Rrhf: Rank responses to align language models with human feedback without tears. *arXiv preprint arXiv:2304.05302*, 2023.

Mohd Zaki, NM Krishnan, et al. Mascqa: A question answering dataset for investigating materials science knowledge of large language models. *arXiv preprint arXiv:2308.09115*, 2023.

Junkai Zhang, Jingru Gan, Xiaoxuan Wang, Zian Jia, Changquan Gu, Jianpeng Chen, Yanqiao Zhu, Mingyu Derek Ma, Dawei Zhou, Ling Li, et al. Matscibench: Benchmarking the reasoning ability of large language models in materials science. *arXiv preprint arXiv:2510.12171*, 2025.

Mengxi Zhang, Wenhao Wu, Yu Lu, Yuxin Song, Kang Rong, Huanjin Yao, Jianbo Zhao, Fanglong Liu, Haocheng Feng, Jingdong Wang, and Yifan Sun. Automated multi-level preference for mllms. In *Advances in Neural Information Processing Systems*, volume 37, pp. 26171–26194, 2024.

Yao Zhao, Rishabh Joshi, Tianqi Liu, Misha Khalman, Mohammad Saleh, and Peter J Liu. Slic-hf: Sequence likelihood calibration with human feedback. *arXiv preprint arXiv:2305.10425*, 2023.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36:55006–55021, 2023.

Zhanhui Zhou, Jie Liu, Jing Shao, Xiangyu Yue, Chao Yang, Wanli Ouyang, and Yu Qiao. Beyond one-preference-fits-all alignment: Multi-objective direct preference optimization. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 10586–10613, 2024.

A Dataset Construction and Rubric Details

A.1 Construction Pipeline and Provenance

The underlying resource is constructed as a provenance-aware QA-score dataset rather than a collection of isolated question–answer pairs. The pipeline contains six stages. *Scientific record governance* normalizes heterogeneous literature-derived records, units, terminology, and composition expressions while retaining raw/source fields for traceability. *Retrieval-corpus construction* converts governed records into structured retrieval entries with material identity, composition, processing, microstructure, property, testing-condition, and application fields. *Scenario modeling* instantiates type-specific semantic slots before question generation. *Retrieval-assisted construction* uses retrieved records to seed scenarios and provide compact evidence. *Multi-answer generation* obtains candidate answers from heterogeneous model sources. Finally, *two-stage rubric scoring* produces total and dimension-level score profiles used by the preference experiments.

The core data unit can be written as

$$p_i = \langle S_i, E_i, C_i, Q_i, A_i, R_i \rangle, \tag{9}$$

where S_i contains source identifiers, E_i retrieved evidence and evidence identifiers, C_i the structured scenario, Q_i the question, A_i candidate answers with source labels, and R_i total/dimension-level score profiles. This representation preserves a lightweight chain from source records through scenarios and evidence to questions, answers, and scores. The same scenario acts as the semantic backbone across question generation, grounded answer generation, and rubric-based scoring.

The resource contains 56,515 retrieval-corpus records, 4,000 questions, 16,000 candidate answers, and 32,000 score records. There are 1,000 questions for each of the four reasoning types. Retrieval is used as construction-time grounding: during question construction it seeds scenario fields, during answer generation compact evidence is injected into prompts, and during scoring the judge receives the question, scenario, candidate answer, and rubric rather than the full raw record. The experiments in this paper therefore do not treat retrieval accuracy as a separate benchmark.

Table 6: Dataset construction statistics.

Item	Count
Retrieval-corpus records	56,515
Questions	4,000
Candidate answers	16,000
Score records	32,000
Reasoning types	4
Questions per type	1,000
Answers per question	4

A.2 Scenario Schemas and Scoring Rubrics

Table 7: Reasoning types and representative scenario fields.

Type	Reasoning target	Representative scenario fields
Causal	Explain whether a material or process variable affects a target property.	Cause, effect, direction, composition, processing condition, property target, application context.
What-if	Propose feasible changes under altered design constraints.	Baseline material, design objective, proposed change, constraints, application domain.
Pathway	Explain a process–microstructure–mechanism–property chain.	Processing, microstructure, mechanism, property, composition, application context.
Mitigation	Diagnose a failure and propose verification and remedy.	Alloy, failure type, symptom, service condition, candidate factor, proposed remedy.

Table 8: Task-specific rubric dimensions and nominal maxima.

Type	Scoring dimensions
Causal	Causal correctness (30), mechanistic depth (30), terminology use (20), logical coherence (20).
What-if	Feasibility (30), technical soundness (30), data support (20), solution diversity (10), clarity (10).
Pathway	Pathway completeness (30), mechanistic clarity (30), logical coherence (20), professional depth (20).
Mitigation	Diagnostic accuracy (30), solution feasibility (30), systematicity (20), operability (20).

Candidate answers are scored under type-specific rubrics and then independently rescored with disagreement-aware fusion. The resulting labels provide both total and dimension-level scores. A sampled subset is additionally used for human consistency checking during dataset construction; this check is distinct from the audited model-output evaluation reported in the main paper. Construction-time validation also checks corpus coverage, answer-source diversity, score consistency across task types, and the stage-specific contribution of retrieval/scenario information. These checks support the use of the resource as an auditable preference study, but they do not constitute independent validation of every generated model answer used in the revised experiments.

B Optimization and Evaluation Details

B.1 Preference-Richness Optimization Matrix

Clean SFT uses high-quality answers and standard autoregressive loss. Complex SFT uses the same autoregressive objective with an orthogonality regularizer and task-specific DoRA experts,

$$\mathcal{L}_{\text{Complex}} = \mathcal{L}_{\text{SFT}} + 0.05\mathcal{L}_{\text{ortho}}. \tag{10}$$

Binary and graded optimization use standard DPO,

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E}\left[\log \sigma\left(\beta\left(\log \frac{\pi_\theta(a^+|q)}{\pi_{\text{ref}}(a^+|q)} - \log \frac{\pi_\theta(a^-|q)}{\pi_{\text{ref}}(a^-|q)}\right)\right)\right], \tag{11}$$

with $\beta = 0.1$. The pair-selection experiments in this study use standard DPO and change only the selected input pairs.

B.2 Training Hyperparameters

Plain SFT uses LoRA rank 8, scale 16, three epochs, learning rate 2×10^{-4}, cosine scheduling with 3% warmup, maximum sequence length 1,024, bf16 precision, per-device batch size one, and gradient accumulation 16. DPO uses learning rate 5×10^{-5}, 4-bit NF4 QLoRA, bf16, maximum sequence length 768, per-device preference-pair batch size 1 with gradient accumulation 16, cosine scheduling, and 3% warmup. The main one-epoch pair-selection runs use an effective batch size of 32, whereas the training-exposure probes use an effective batch size of 16; the corresponding optimizer-update and pair-exposure counts are summarized in Section 4.3 and analyzed in Section 5.4. Iterative training uses learning rate 5×10^{-6}. Experiments use Qwen3.5-9B generators and a Qwen3.5-4B-based frozen ScoreLLM evaluator.

B.3 Evaluation Audit and Metric Scope

The evaluation audit covers the reported historical and current checkpoints. Configurations whose labels did not match their realized training objectives are excluded from the validated method matrix. The retained runs cover pair-selection and ablation runs, SADD variants, and the preference-richness optimization runs. Files are aligned by unique qid, with expected per-type counts 105/104/102/102. The prior parser could map repeated rubric-output blocks to zero; the robust parser takes the first occurrence of each expected rubric dimension and recovers roughly 100 such parser-induced failures. Some earlier files also had qid mismatches, and the Clean-SFT baseline was regenerated because no

compatible cached generation was available. Consequently, changes between pre-audit and audited scores have multiple causes and are not attributed to one error source.

Dataset-original A/B/C/D labels and scores for newly generated answers are produced under different protocols. We therefore use $\bar{S}_{\text{gen}}$ as the central cross-model metric. Reference-relative gap, Surpass%, and Close% are reported only as secondary diagnostics. For fixed checkpoints, paired bootstrap resamples matched per-question differences 10,000 times; these intervals measure question-level variation conditional on the trained checkpoints and sampled outputs, not method-level uncertainty over training seeds.

C Audited Baseline Results

This appendix reports the validated preference-richness optimization matrix and a compact per-type breakdown used only for descriptive interpretation in the main text. Experiment IDs are retained from the original run registry; configurations excluded after implementation audit are omitted from the reported matrix. All rows use the common 413-question evaluation described in Section 4; the original-matrix checkpoints are single trained instances and should not be read as multi-seed method estimates.

C.1 Validated Preference-Richness Matrix

Table 9: Validated preference-richness matrix under the audited common evaluation.

ID	Configuration	$\bar{S}_{\text{gen}}$
M0	Zero-shot	75.6
M1	Clean SFT	86.8
M2	Clean-start binary DPO	92.6
M3	Clean-start graded DPO	92.6
M4	Complex-start binary DPO	91.5
M5	Complex-start graded DPO	92.6
M7	Iteration 1	92.5
M8	Iteration 2	92.4
M9	Iteration 3	92.5
M10	Iterative final	92.4
M11	OrthoAux + DPO	92.3

C.2 Representative Per-Type Breakdown

The per-type means illustrate heterogeneity across causal, what-if, pathway, and mitigation questions, but no type-specific training-seed replication is available for these original-matrix checkpoints. We therefore use the table only to show that the aggregate ordering need not hold identically for each reasoning type.

Table 10: Representative per-type means from the audited preference-richness matrix. These are descriptive single-checkpoint results.

Configuration	Causal	What-if	Pathway	Mitigation
M2: clean binary	96.5	87.3	92.2	94.2
M3: clean graded	96.3	86.6	93.5	93.9
M5: complex graded	96.2	86.7	93.7	93.9

D Additional Pair-Selection Analyses

D.1 Budget Sweep

The budget sweep asks whether the seed-42 score curve changes smoothly as progressively more of the pair pool is retained. Because q25 and q75 scalar/rubric conditions are not replicated across

seeds, these points are treated as descriptive shape probes rather than evidence for an optimal retention fraction.

Table 11: Seed-42 budget sweep. q25 and q75 scalar/rubric results are single-seed descriptive probes; q100 is the full pair pool.

Selection rule	q25	q50	q75	q100
Rubric-abs	92.5	92.8	92.7	92.3
Scalar	92.4	93.0	92.9	92.3
Random	92.5	92.7	92.5	92.3

At q25, random selection has three seeds (92.47, 91.09, 90.86; mean 91.47 ± 0.87), whereas scalar and rubric-absolute q25 are available only for seed 42. The current experiments therefore do not support a multi-seed claim that high-gap selection systematically outperforms random selection at the smallest tested budget.

D.2 Selected-Pair Composition

The following table describes what the q50 selectors actually retain. Gap bins use the dataset-original scalar total, whereas the conflict column asks whether at least one rubric dimension favors the rejected answer. These quantities characterize selection behavior and should not be interpreted as causal performance contributions.

Table 12: Selected-pair properties at nominal q50. Gap bins use the original scalar total score. "Conflict" means at least one rubric dimension favors the rejected answer.

Rule	$\Delta S \geq 25$	$15 \leq \Delta S < 25$	$\Delta S < 15$	Conflict
Rubric-abs	10.5%	49.0%	40.4%	3.9%
Rubric-signed	10.5%	49.0%	40.6%	1.9%
Consensus	10.4%	48.4%	41.3%	1.4%
Scalar	10.3%	48.0%	41.7%	2.5%
Random	5.4%	24.7%	69.9%	12.0%
Bottom	0.0%	0.0%	100.0%	20.0%
Top–bottom only	20.1%	59.3%	20.6%	2.9%

Rubric-absolute versus scalar q50 has Jaccard overlap 0.943. Random-versus-gap overlap is approximately 0.34 and bottom-versus-gap overlap approximately 0.02.

D.3 Sensitivity to the Rubric Analysis Unit

The main-text audit treats individual scored answers as observations. This can retain question-level shared variation: for example, an easier question may receive relatively high scores across several rubric dimensions. To test whether the qualitative redundancy pattern depends on this analysis unit, we recompute the same statistics on the six within-question rubric-gap vectors $\Delta\mathbf{s} = \mathbf{s}(a^+) - \mathbf{s}(a^-)$.

For a matched sensitivity comparison, Table 13 uses the full 4,000-question resource for both units. Correlations and PCA are computed with the same Pearson, dimension-standardization, and 95% effective-dimensionality conventions as in the main-text audit.

Table 13: Sensitivity of rubric redundancy statistics to the analysis unit. Answer-level and pair-gap results use the same full 4,000-question resource.

	Answer level			Pair gap		
Type	Mean r	PC1	Dim.	Mean r	PC1	Dim.
Causal	.829	87.4%	2	.713	78.9%	3
What-if	.583	67.6%	4	.437	56.1%	4
Pathway	.954	96.6%	1	.894	92.1%	2
Mitigation	.900	92.5%	2	.771	82.9%	3

Pair-gap analysis lowers the absolute correlation and PC1 variance for all four reasoning types, consistent with part of the answer-level covariance reflecting shared question-level variation. The qualitative heterogeneity nevertheless remains: pathway is still strongly correlated ($r = .894$), whereas what-if remains the least correlated type ($r = .437$). Thus, the answer-level audit should be interpreted as a descriptive property of the scored resource rather than as direct evidence that all rubric dimensions are redundant at the preference-pair level.

D.4 Leave-One-Pair-Position Diagnostic

The leave-one-pair-position runs hold retained-example count and nominal optimizer steps fixed across the six omissions. They are intended to test whether pair positions are interchangeable, not to rank the intrinsic semantic value of individual pair types.

Table 14: Exploratory seed-42 leave-one-pair-position-out runs. Each condition keeps 15,965 pairs and approximately 998 steps.

Omitted pair	$\bar{S}_{\text{gen}}$	vs. full 92.31
(A, B)	90.01	-2.30
(C, D)	91.66	-0.65
(A, C)	91.92	-0.39
(A, D)	92.07	-0.24
(B, D)	92.43	+0.12
(B, C)	92.72	+0.41

These runs are internally matched in example count and nominal steps, but removing one pair position also changes the chosen/rejected role frequency of its endpoints. They therefore diagnose pair-position sensitivity without identifying the intrinsic semantic utility of a pair type.

D.5 Fixed-Checkpoint Bootstrap Diagnostics

These paired-bootstrap comparisons quantify variation over matched test questions for particular trained checkpoints. They complement, but do not replace, the multi-seed evidence in the main text.

Table 15: Selected seed-42 paired-bootstrap diagnostics. Intervals resample the 413 matched questions conditional on fixed checkpoints and do not include training-seed uncertainty.

Comparison	Mean difference	95% CI	p
Full truncated-exposure vs. full one epoch	+0.54	$[+0.33, +0.75]$	$< .001$
Consensus q50 two epochs vs. full	+0.51	$[+0.30, +0.72]$	$< .001$
Scalar q50 one epoch vs. full	+0.73	$[+0.50, +0.95]$	$< .001$
Scalar q50 two epochs vs. full	-0.60	$[-1.01, -0.24]$	.003
Scalar q50 vs. rubric-abs q50	+0.21	$[+0.05, +0.37]$	.010
Consensus q50 vs. scalar q50	-0.13	$[-0.28, +0.04]$	.133

These tests are supplementary because they do not model training-seed variation and the collection of pairwise comparisons is not corrected for multiplicity.

E Exploratory Direct Dimension-Aware Optimization

We additionally evaluated an exploratory objective that directly optimizes rubric-specific quality margins predicted from pooled hidden representations. Unlike the main pair-selection experiments, which use rubric scores only to select training pairs and then apply standard DPO, this objective introduces dimension-level supervision directly into optimization:

$$\mathcal{L}_{\text{dim}} = -\mathbb{E}\left[\sum_{k=1}^{K} w_k \Delta s_k \log \sigma\left(\beta \frac{\hat{q}_k(a^+) - \hat{q}_k(a^-)}{\tau_k}\right)\right]. \tag{12}$$

Here $w_k = M_k / \sum_j M_j$, $\Delta s_k = s_k(a^+) - s_k(a^-)$, and $\hat{q}_k(a)$ is the output of the dimension-specific quality head applied to the pooled projected hidden representation of answer a. We use $\beta = 0.1$ and $\tau_k = \tau_0 \sigma_k / \bar{\sigma}$ with $\tau_0 = 0.07$, where σ_k is the training-set standard deviation of rubric dimension k. The dimension-specific quality heads were pretrained on rubric-level pairwise ordering before generator optimization. Under otherwise matched seed-42 conditions, complex-start graded DPO obtains 92.63, whereas the direct dimension-aware objective obtains 75.46.

Table 16: Exploratory seed-42 direct dimension-aware optimization.

Configuration	$\bar{S}_{\text{gen}}$
Complex-start graded DPO	92.63
Direct dimension-aware objective	75.46

Training diagnostics indicate that the dimension-margin objective itself optimizes successfully while generated-answer quality deteriorates. Because this experiment uses a single seed and a particular quality-head architecture, we treat it as an exploratory failure analysis rather than evidence that direct dimension-aware optimization is generally ineffective.

F Secondary Semantics-Aware Dimension Decomposition Experiments

We additionally study Semantics-Aware Dimension Decomposition (SADD), an exploratory rubric-weighting strategy that diagnoses dimension-level scoring structure before preference-pair selection. SADD is included as a secondary analysis rather than as a primary method contribution because all three variants are evaluated with a single training seed.

Dimension-reliability diagnosis. SADD compares the dimension-correlation structures produced by two separately trained scoring-model runs, denoted ScoreModel410 and ScoreModel412. Both use the same Qwen3.5-4B backbone, LoRA adaptation, and scoring annotations, but arise from different training runs. For correlation matrices Σ_A and Σ_B, SADD estimates a shared-correlation component from the agreement between the two judges. For dimension k, a diagnostic weight is then constructed as

$$w_k \propto \text{signal_ratio}_k \cdot \text{uniqueness}_k \cdot \text{base_weight}_k, \tag{13}$$

where signal_ratio_k measures the fraction of the dimension's correlation structure shared by the two scoring models, uniqueness_k discounts dimensions that are strongly correlated with the remaining shared dimensions, and $\text{base_weight}_k = M_k/100$ retains the task-specific rubric scale. The normalized weights are used to construct a weighted rubric-gap score for q50 preference-pair selection.

We compare three variants under the same nominal q50 data budget and one-epoch DPO training: SADD-DPO uses the diagnostic weights above, SADD-DPO-equal assigns uniform weights $1/K$, and SADD-DPO-expert uses the rubric maxima $M_k/100$ as expert-defined weights.

Table 17: Secondary SADD rubric-weighting experiments (seed 42). All three variants use the same nominal q50 data budget and one training epoch.

Method	$\bar{S}_{\text{gen}}$
SADD-DPO-equal	92.2
SADD-DPO (triangulation weighting)	91.8
SADD-DPO-expert	91.0

Among these single-seed runs, equal weighting has the highest observed score (92.2), followed by triangulation weighting (91.8) and expert rubric weighting (91.0). These are single-seed results and therefore do not establish a reliable ordering among weighting strategies. Moreover, the two scoring models share the same base-model family and training annotations, so their shared correlation cannot be interpreted as an independent estimate of ground-truth dimension reliability. We therefore treat SADD as an exploratory diagnostic showing that more elaborate rubric weighting does not automatically translate into better alignment, rather than as evidence for a primary method contribution.