

The Variance of a Judge: Zero-Variance Rewards by Deterministic Regression

Kabir Murjani¹ Parth Vyas¹ Akshita Abrol^{2*} Rajesh Gupta³

Sudeep Tanwar³ Zhengkui Wang²

¹Nirma University ²Singapore Institute of Technology ³Marwadi University

Abstract

Large Language Model (LLM) judges score responses by generating rating tokens, so their rewards carry epistemic, generation-induced uncertainty from the scoring mechanism itself, not only genuine disagreement about quality. Neighbour-Oriented Reward Anchoring (NORA) replaces generative scoring with retrieval-augmented regression: a small MLP, the Optimised Response Latent Inference (ORLI) judge, scores each query–response pair over a frozen embedding manifold anchored to human-rated neighbours. Because no language model generates the score, the reward is deterministic and its generation-induced variance is zero by construction. This relocates the error rather than removing it: the residual (MAE = 0.736) is dominated by systematic bias, not run-to-run noise. ORLI is still a learned reward, so NORA curbs over-optimisation with a retrieval-guided curriculum and a contrastive penalty. Gains are asymmetric and largest where quality is lowest, on Recipe4U and HelpSteer2 alike, and a blinded human evaluation reproduces the pattern.

1 Introduction

A learned reward is only as useful as our understanding of where its errors come from. In reinforcement learning from human feedback (RLHF), that reward increasingly comes from a generative “LLM-as-a-Judge” (Ouyang et al., 2022), which produces a rating token through the same autoregressive machinery it uses to generate text. This entangles two sources of uncertainty in the reward: *aleatoric* disagreement among annotators about

how good a response is, and *epistemic, generation-induced* variance that originates in the scoring mechanism rather than the response. LLM judges exhibit the latter in familiar forms: they overrate plausible-but-empty responses (Zheng et al., 2023), favour their own generations (Panickssery et al., 2024), and shift verdicts with prompt phrasing and sampling noise. Throughout, we characterise a scorer by its generation-induced (epistemic) variance (Definition 1), the extent to which its score moves while the response does not. This is a narrower property than “hallucination” as that term is generally used, and it bears on variance rather than correctness: a deterministic scorer can still be systematically wrong, and ours is. Traditional scalar reward models avoid token generation but overfit annotation artefacts (Gao et al., 2023) and need annotation at a scale most groups cannot afford (Ouyang et al., 2022).

We propose Neighbour-Oriented Reward Anchoring (NORA) (Figure 1). Its scoring component, Optimised Response Latent Inference (ORLI), is a compact MLP that interpolates a quality score by regression over a frozen embedding space, using nearby human-labelled examples from a rating-partitioned memory. Because no language model ever generates a rating token, the reward’s generation-induced variance is eliminated by construction (identical inputs yield identical scores), leaving approximation bias and the aleatoric uncertainty of the annotations as the remaining error terms. This is a statement about variance, not about correctness: ORLI makes no fewer errors than its bias permits, it simply makes the same error every time. We still call ORLI a judge, since it fills the role an LLM judge would, and describe NORA as free of generative scoring rather than judge-free. What is removed is autoregressive generation from the scoring path, not the judge.

Zeroing V_{gen} does not by itself prevent over-optimisation: like any learned reward, ORLI can

* Author emails: Kabir Murjani, 23bee064@nirmauni.ac.in; Parth Vyas, 24mce026@nirmauni.ac.in. Corresponding Author Email: akshita.abrol@singaporetech.edu.sg
rajesh.gupta@marwadieducation.edu.in
sudeep.tanwar@marwadieducation.edu.in
zhengkui.wang@singaporetech.edu.sg

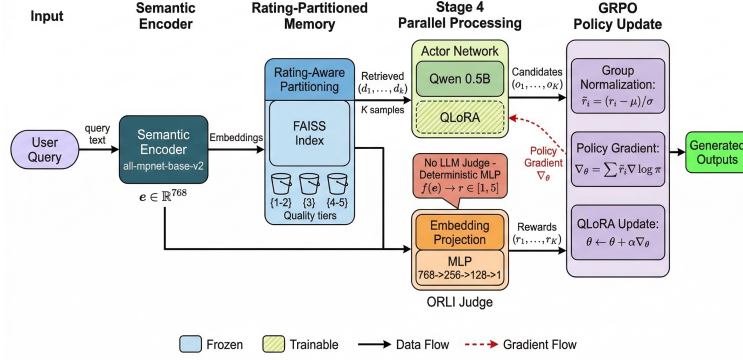


Figure 1: NORA framework overview: a frozen semantic encoder produces embeddings that feed into rating-partitioned FAISS indices for retrieval and into the ORLI judge (a lightweight MLP) for quality scoring. The actor network (Qwen2-0.5B with QLoRA) generates candidate responses, and GRPO updates the policy using group-normalised advantages from ORLI rewards. No language model participates in scoring.

be exploited (Gao et al., 2023) until coherence collapses (Table 4). NORA therefore *mitigates over-optimisation through retrieval-guided curriculum learning*: a curriculum (Section 3.2) anneals the reward from similarity with retrieved gold responses toward ORLI scores, and a contrastive penalty (Section 3.3) pushes the lowest-rated generations away from known poor examples. Policy updates use Group Relative Policy Optimisation (GRPO) (Shao et al., 2024), which normalises advantages within response groups, needs no value network, and, unlike Direct Preference Optimisation (DPO) (Rafailov et al., 2023), consumes pointwise 1–5 scores directly.

We evaluate on two public dialogue datasets, Recipe4U (tutoring; Han et al. 2024) and Help-Steer2 (assistant responses; Wang et al. 2024). With a 0.5B actor, NORA raises 1–2 star responses by up to +0.77 ORLI points while leaving 4–5 star outputs largely unchanged, confirmed by a blinded human evaluation (Section 5.5). One caveat should be noted at the outset: ORLI is both the training reward and the primary automatic metric, a circularity we treat as a limitation and partly address with the human study.

The main contributions are:

1. A formalisation of the generation-side component of reward error as generation-induced variance V_{gen} (Definition 1), with an error-budget decomposition (Eq. 3) separating it from aleatoric noise and approximation bias. ORLI, a retrieval-augmented MLP that replaces autoregressive reward generation with deterministic embedding regression, drives

V_{gen} to zero by construction (Proposition 1), leaving approximation bias and aleatoric annotation uncertainty. We measure the residual with the judge’s held-out MAE and split it per tier into signed bias and magnitude (Appendix N), finding bias dominant. Zeroing V_{gen} is thus a reallocation of the error budget, not a reduction of it.

2. A retrieval-guided curriculum reward schedule that mitigates reward over-optimisation by annealing from similarity-based imitation to quality optimisation, paired with contrastive regularisation for the lowest-rated tier.
3. Empirical evidence that NORA improves every rating tier with the largest gains where quality is lowest, a pattern that holds across two datasets, multiple seeds and group sizes, and is confirmed by blinded human evaluation.

2 Related work

Reward models, judges, and their uncertainty.

Ouyang et al. (2022) established the RLHF paradigm, in which a reward model trained on human comparisons guides policy optimisation. Generative LLM judges have since become a common reward source, but they inherit the failure modes of autoregressive generation: they reward fluency over substance and are prone to reward hacking (Skalse et al., 2022; Zheng et al., 2023), and they favour their own outputs (Panickssery et al., 2024). In the language of uncertainty these are epistemic, generation-induced errors: the score depends on

how it is produced, not only on what is scored. Gao et al. (2023) showed that reward models struggle to guide policies past annotator quality and that over-optimisation eventually degrades it, while reward-model ensembles mitigate this by exposing epistemic uncertainty in the reward (Coste et al., 2024). Parameter-efficient tuning (Hu et al., 2022; Dettmers et al., 2023) lowers cost but leaves these reward pathologies untouched.

Synthetic feedback and critic-free optimisation. RLAIF (Lee et al., 2024), Constitutional AI (Bai et al., 2022), and ALMoST (Kim et al., 2023) bootstrap preferences from teacher models, but inherit the teacher’s biases and the same self-preference effects as direct judges. NORA instead reads scores off human data by nearest-neighbour interpolation (Khandelwal et al., 2020) over a frozen manifold, so no model generates the reward. For the policy update, DPO (Rafailov et al., 2023) drops the explicit reward model but is limited to pairwise preferences and cannot consume pointwise ratings; GRPO (Shao et al., 2024) removes the value network by normalising advantages within response groups and underpins strong critic-free RL (DeepSeek-AI, 2025). We pair GRPO with a deterministic, retrieval-augmented scorer in place of a generative or scalar reward model.

Relation to ordinal reward models. Proposition 1 is not unique to NORA: any deterministic scalar or ordinal reward model $g_\psi(x, y)$ trained on human judgements and evaluated with fixed weights also has $V_{\text{gen}} = 0$. We therefore do not claim determinism as a contribution over that class. Determinism is claimed only against generative judges, which remain the dominant reward source in current RLHF practice and which do pay the V_{gen} term. Stated plainly: Proposition 1 separates NORA from LLM-as-a-judge, not from reward models.

ORLI nevertheless differs from an ordinal reward model in one substantive respect. Both map a pair to a scalar and both are fitted to human labels, but they differ in where the inductive bias for interpolation lives. An ordinal reward model interpolates through the generalisation of a trained network over a representation it is itself free to reshape, so its behaviour away from the training data is governed by learned parameters and is correspondingly hard to inspect. ORLI interpolates over a frozen manifold that neither the judge nor the policy can deform, with a rating-partitioned

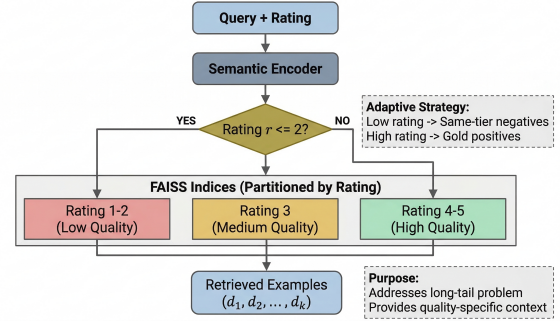


Figure 2: Rating-aware retrieval strategy, in which low-rated queries ($r \leq 2$) retrieve contrastive negatives from the same tier, while higher-rated queries retrieve gold exemplars ($r \geq 4$) to guide imitation. The FAISS indices are partitioned by rating bucket to enable tier-specific nearest-neighbour search.

neighbour structure (Figure 2) that makes the support of any score explicit: the neighbours responsible for a prediction can be retrieved and read. Three consequences follow, and we flag which are demonstrated here and which are not. (i) The same frozen manifold serves retrieval, the curriculum’s similarity term, and scoring, so the reward and the imitation anchor are defined in one geometry rather than two, which is what the curriculum ablation (Table 4) exercises. (ii) The memory is editable without retraining: adding or removing rated examples changes the neighbour sets and hence the scores directly, whereas an ordinal reward model requires a fitting step. (iii) Scores are attributable to specific human-rated exemplars. Claims (ii) and (iii) are properties of the construction that we have not measured, and we do not rest any empirical result on them.

This is a positioning argument rather than an empirical one. We report no head-to-head comparison against a trained ordinal reward model or an LLM judge, so the practical advantage of nearest-neighbour interpolation over learned generalisation remains open; Section 5.4 sets out the required comparison and Appendix O records it as a limitation.

3 Methodology

We consider a dataset $\mathcal{D} = \{(x_i, y_i, r_i)\}_{i=1}^N$, where each query x_i is paired with a response y_i and a human quality rating $r_i \in \{1, \dots, 5\}$. The goal is to learn a policy π_θ whose generated response $\hat{y} \sim \pi_\theta(\cdot | x)$ attains higher quality than the original, particularly for low-rated examples.

Where NORA applies. NORA targets a specific data regime, and we state its assumptions before describing the mechanism. It expects dense, point-wise human-rated dialogue data: enough rated examples that a query’s k nearest neighbours in embedding space are genuinely near it, ratings attached to individual responses rather than pair-wise preferences, and a quality notion that semantic similarity can track. Both our datasets meet this: Recipe4U with 1,913 rated tutoring dialogues and HelpSteer2 with 20,324 rated responses, each on a 1–5 scale over open-ended natural-language answers.

It is correspondingly unsuited to three regimes. In sparse-annotation settings the neighbour sets become too sparse to constrain the interpolation, which degrades toward the MLP’s global prior; we did not vary training-set size, so we cannot say where that threshold sits. In verifiable domains such as code or mathematics, correctness is not a smooth function of embedding proximity, since one flipped operator moves the answer from right to wrong while barely moving the vector, so execution- or proof-based rewards are preferable, and ORLI could at best supplement them. Where only pairwise preferences exist, there are no point-wise targets to regress onto, which is the setting DPO (Rafailov et al., 2023) addresses instead.

The imitation ceiling bounds the achievable gain.

Standard reward models face an *imitation ceiling*: since R_ϕ approximates the annotators’ rating distribution, the policy cannot reliably surpass the best training examples (Gao et al., 2023). NORA eases this through two mechanisms: the embedding manifold supplies a continuous quality signal that interpolates between observed ratings, and the curriculum schedule (Section 3.2) controls how aggressively the policy exploits it, preventing collapse toward high-rated demonstrations early while allowing quality-driven optimisation later.

It eases the ceiling without lifting it, and this bounds the achievable gain in a manner that plausibly accounts for the observed asymmetry. ORLI’s scores are interpolations between observed human ratings, so the reward carries no information about quality above the best-rated neighbourhoods in the memory: the signal is informative about the gap between a poor response and a good one, and nearly flat between a good response and an excellent one. Two consequences follow from this constraint. First, the gain available at a tier is roughly

the headroom between it and the local gold ceiling, which is large at 1-star and close to zero at 5-star. This is the shape observed empirically (Table 2), and it persists net of a matched untrained-policy null (Appendix N). Our results should therefore be read as NORA efficiently closing the distance to the rated ceiling, not as evidence of improvement beyond it. Second, the ceiling interacts with the judge’s per-tier bias: ORLI underrates the top of the scale by -1.30 at 5-star (Table 11), so even genuine gains there would be partly invisible to the reward that is supposed to reinforce them. Surpassing annotator quality would need a signal ORLI does not have, whether verifiable rewards, stronger annotations, or active collection at the frontier.

3.1 System architecture and ORLI scoring

NORA has five components: (i) a frozen sentence encoder E ; (ii) a rating-partitioned FAISS memory built over the training split; (iii) the ORLI judge f_ϕ ; (iv) a QLoRA actor π_θ ; and (v) a rating-stratified prioritised replay buffer feeding GRPO updates. All five share an embedding space produced by a frozen all-mpnet-base-v2 encoder (Reimers and Gurevych, 2019): every query–response pair passes through this 768-dimensional encoder first, so retrieval and scoring share one representational basis throughout training. This couples our results to the chosen encoder, as the Limitations section discusses.

The design choice that distinguishes NORA from reward-model and LLM-judge alternatives is replacing autoregressive score generation with deterministic regression over this frozen manifold, performed by the Optimised Response Latent Inference (ORLI) judge. Given a pair (x, y) , we encode the joint input "Query: x Response: y " into a vector $e \in \mathbb{R}^{768}$ and compute a scalar quality score via a three-layer MLP:

$$f_\phi(e) = W_3 \cdot \text{GELU}(\text{LN}(W_2 \cdot \text{GELU}(\text{LN}(W_1 \cdot e)))) \quad (1)$$

with dropout ($p=0.2$ then $p=0.1$) to regularise the small parameter count. We pretrain the judge for 15 epochs on training-split labels using Huber loss ($\delta=1.0$), which down-weights gradient contributions from rating outliers and stabilises convergence. On a held-out validation split the pretrained judge achieves MAE = 0.736, roughly three-quarters of a rating point on the 1–5 scale. Because the scorer is deterministic, this error carries no run-to-run variance. It should not, however,

be read as the aleatoric floor: the per-tier decomposition in Appendix N shows that the residual is largely systematic bias, with the judge overrating low-rated responses and underrating high-rated ones. Determinism is also a precondition for calibration: a stochastic judge spreads scores over a single input, which no fixed calibration map can absorb. After pretraining, the judge is frozen and receives no further gradient updates during policy optimisation. Figure 5 (Appendix I) provides an architectural diagram.

Greedy decoding does not achieve the same effect. It fixes the sampling draw, but the rating token is still produced through sequential vocabulary predictions shaped by surface cues, self-preference bias (Panickssery et al., 2024), and prompt sensitivity (Zheng et al., 2023), leaving the prompt- and ordering-induced variation in Z untouched.

We can now state this claim precisely. Let a scorer assign to (x, y) a score $S(x, y; Z)$, where the nuisance variable Z collects generation-side stochasticity: sampling seed, temperature, prompt template, and option ordering.

Definition 1 (Generation-induced reward variance). *The generation-induced (epistemic) variance of a scorer at (x, y) is*

$$V_{\text{gen}}(x, y) = \text{Var}_Z [S(x, y; Z)]. \quad (2)$$

A scorer carries epistemic, generation-side uncertainty on (x, y) to the extent that $V_{\text{gen}}(x, y) > 0$: the score varies while the response does not. We name only this quantity; it is one component of reward error, not a synonym for judge failure in general.

Proposition 1 (ORLI eliminates generation-induced variance). *ORLI scores through $S(x, y) = f_\phi(E(x, y))$, where the encoder E and the MLP f_ϕ are both frozen, deterministic, and independent of Z . Then $S(x, y; Z)$ is constant in Z , so $V_{\text{gen}}(x, y) = 0$ for every (x, y) .*

The proof is immediate: a deterministic function of a Z -independent argument has zero variance over Z . Writing r^* for the mean human rating and R for a fresh label with $\mathbb{E}[R] = r^*$ and aleatoric variance σ_{alea}^2 , the expected squared error decomposes into three terms,

$$\mathbb{E}_{Z,R}[(S - R)^2] = V_{\text{gen}} + b^2 + \sigma_{\text{alea}}^2, \quad (3)$$

the generation-induced variance, the approximation bias $b = \mathbb{E}_Z[S] - r^*$, and the aleatoric annotation

variance. A generative judge pays all three; by Proposition 1 ORLI zeroes the first, so its residual error (empirically MAE = 0.736) is governed by the aleatoric term and the MLP’s bias, not by the scoring dynamics.

This is a reallocation of the error budget, not a reduction of it. Zeroing V_{gen} says nothing about the size of b , and a deterministic scorer with large bias can be worse than a noisy unbiased one; Appendix N indicates b is in fact ORLI’s dominant term. We also do not measure V_{gen} for a generative judge here, so the size of the term being removed is taken from the literature (Zheng et al., 2023; Panickssery et al., 2024) rather than established by us.

Retrieval feeds into ORLI through a rating-partitioned FAISS index (Johnson et al., 2021): the training set is stored in partitions keyed by rating bucket (Figure 2), and routing depends on the query’s tier. Low-rated queries ($r \leq 2$) retrieve k same-tier neighbours as contrastive negatives (Section 3.3), while queries rated 3 or above retrieve gold references ($r \geq 4$) as imitation targets for the curriculum reward (Section 3.2).

The actor is Qwen2-0.5B-Instruct with QLoRA (Dettmers et al., 2023) (rank-8, 4-bit NF4, $\alpha=32$), $\sim 0.5\text{M}$ trainable parameters (0.1%); its compact footprint preserves GPU memory for the FAISS indices and buffers. Following Schaul et al. (2016), a prioritised experience replay (PER) buffer stratified by rating gives higher priority to lower-rated examples, concentrating optimisation where the quality gap is widest (settings in Appendix F).

3.2 Curriculum reward learning

Direct optimisation on quality scores causes mode collapse. We address this with a curriculum that transitions from imitation to optimisation (Bengio et al., 2009); rationale and parameters are in Appendix B. The composite reward at step t is:

$$R(x, \hat{y}; t) = \alpha(t) \cdot R_{\text{sim}}(x, \hat{y}) + \beta(t) \cdot R_{\text{ORLI}}(x, \hat{y}) \quad (4)$$

where R_{sim} measures cosine similarity to retrieved gold responses. The weighting coefficients anneal linearly:

$$\alpha(t) = 1.0 - 0.8 \cdot (t/T) \quad (5)$$

$$\beta(t) = 0.1 + 0.7 \cdot (t/T) \quad (6)$$

Early training ($\alpha \approx 1, \beta \approx 0.1$) anchors the policy to human demonstrations; later training ($\alpha \approx 0.2$,

$\beta \approx 0.8$) shifts toward quality, but the similarity signal never vanishes, preventing unbounded drift.

3.3 Contrastive regularisation

For low-rated samples ($r \leq 2$), we add a contrastive objective that penalises similarity to known poor responses (derivation in Appendix C). Rating-aware retrieval provides negative examples Y^- from the same quality tier:

$$\mathcal{L}_{\text{contrastive}} = \frac{1}{|Y^-|} \sum_{y_j^- \in Y^-} \text{sim}(E(\hat{y}), E(y_j^-)) \quad (7)$$

Minimising this loss reduces the mean cosine similarity between $\hat{y}$ and the retrieved negatives, pushing the generation away from low-quality patterns (Gao et al., 2021). Since observed similarities are positive, $\mathcal{L}_{\text{contrastive}}$ enters the loss with weight $\lambda_c = 0.3$ and no leading minus sign, so gradient descent drives the embedding away from the negatives (Appendix C). It activates only for low-rated samples.

3.4 Policy optimisation via GRPO

We adopt GRPO (Shao et al., 2024) for its critic-free design (derivation in Appendix D). For each query we sample K candidates at temperatures drawn uniformly from $[0.7, 1.8]$; the high upper bound encourages diversity. No filtering is applied, since group normalisation down-weights degenerate low-scoring candidates. The group-normalised advantage is:

$$A_i = \frac{R_i - \mu_g}{\sigma_g + \epsilon} \quad (8)$$

where μ_g and σ_g are the group mean and standard deviation. Because queries are drawn from the replay buffer non-uniformly, the policy gradient is corrected by the importance-sampling weight $w(x)$ attached to the sampled query (Appendix D):

$$\mathcal{L}_{\text{policy}} = -w(x) \sum_{j=1}^K A_j \cdot \log \pi_{\theta}(\hat{y}_j | x) \quad (9)$$

The total objective combines policy and contrastive terms: $\mathcal{L} = \mathcal{L}_{\text{policy}} + \lambda_c \cdot \mathcal{L}_{\text{contrastive}}$. Algorithm 1 (Appendix E) summarises the complete training procedure. Eliminating PPO’s value network frees memory for FAISS indices and experience buffers, enabling training on a single on a cloud-hosted GPU with 80GB of memory in 7 h ($K=16$) or 14 h ($K=32$).

4 Experimental setup

4.1 Datasets

We evaluate on two publicly available preference datasets. Full statistics appear in Appendix G.

Recipe4U. Recipe4U (Han et al., 2024) has 1,913 student–tutor dialogues with 1–5 star ratings. The distribution is skewed: 43% receive 5 stars, only 7% are rated 1–2, motivating rating-aware retrieval. We use an 80/20 stratified split.

HelpSteer2. HelpSteer2 (Wang et al., 2024) provides 20,324 model-generated responses rated for helpfulness on a 0–4 scale, which we map to 1–5. We use an 80/20 stratified split and the helpfulness axis only; its other axes and their joint optimisation are left for future work.

4.2 Baselines

Gold responses. The original human responses from each test set. “Gold” denotes the original paired response, not a quality label (many received low ratings); all Δ values are computed against them.

Supervised fine-tuning (SFT). Qwen2-0.5B-Instruct fine-tuned on high-rated responses ($r \geq 4$) using cross-entropy loss, isolating pure imitation learning without RL.

4.3 Evaluation metrics

ORLI score. The quality score from a frozen ORLI judge, trained only on training-split labels and with no access to test-set ratings. ORLI is also the training reward; this dual role is a limitation we discuss later. Scores range from 1 to 5.

Quality delta (Δ). The improvement of generated responses over their gold counterparts:

$$\Delta = R_{\text{ORLI}}(\hat{y}) - R_{\text{ORLI}}(y_{\text{gold}}) \quad (10)$$

Per-rating delta (Δ_r). We disaggregate Δ by original rating bucket:

$$\Delta_r = \frac{1}{|\mathcal{S}_r|} \sum_{i \in \mathcal{S}_r} (R_{\text{ORLI}}(\hat{y}_i) - R_{\text{ORLI}}(y_i)) \quad (11)$$

where $\mathcal{S}_r = \{i : r_i = r\}$.

Semantic similarity. Cosine similarity between generated and gold response embeddings:

$$\text{Sim}(\hat{y}, y) = \frac{E(\hat{y}) \cdot E(y)}{\|E(\hat{y})\| \|E(y)\|} \quad (12)$$

4.4 Implementation details

Retrieval uses $k=3$ neighbours, and policy training runs 3 epochs $\times$ 80 batches with AdamW ($\text{lr}=2 \times 10^{-5}$). Complete hyperparameters are listed in Appendix F.

5 Results and analysis

We evaluate NORA on Recipe4U ($N_{\text{test}}=382$) and HelpSteer2 ($N_{\text{test}}=4,063$). For each dataset we report $K=16$ with two random seeds (42, 123) and $K=32$ with seed 42. Statistical significance tests for the Recipe4U $K=32$ configuration appear in Appendix H.

All improvements below are measured against gold responses (the original human outputs), not against alternative alignment methods; the SFT comparison characterises the qualitative pattern rather than claiming superiority.

5.1 Overall quality gains and per-rating asymmetry

Table 1 presents the primary metrics. Across both datasets and all configurations, NORA generates responses that ORLI rates higher than the human originals, with overall gains from +0.091 points on HelpSteer2 ($K=16$) to +0.152 on Recipe4U ($K=32$). Cosine similarity to gold stays between 0.55 and 0.60, so outputs remain related to the originals without collapsing into paraphrase. Raising the group size from $K=16$ to $K=32$ gives a consistent but diminishing boost (Appendix K).

Dataset	Method	K	Seed	ORLI	Δ	Sim
Recipe4U	Gold Responses	–	–	4.333	–	1.000
	NORA	16	42	4.446	+0.112	0.591
	NORA	16	123	4.466	+0.133	0.594
	NORA	32	42	4.485	+0.152	0.603
HelpSteer2	Gold Responses	–	–	3.870	–	1.000
	NORA	16	42	3.961	+0.091	0.548
	NORA	16	123	3.962	+0.092	0.551
	NORA	32	42	3.977	+0.107	0.559

Table 1: Quality improvements across both datasets, where Δ denotes improvement over gold responses as measured by ORLI; Sim denotes cosine similarity to gold.

The aggregate figures conceal an asymmetry that emerges only once the results are disaggregated by original rating. Table 2 breaks the improvements down by quality tier (plotted in Figure 3). On Recipe4U, 1-star examples improve by +1.17 ORLI points on average, roughly a full rating level, while 5-star examples shift by +0.10. HelpSteer2 exhibits the same gradient (+0.62 for 1-star, +0.03

for 5-star) despite a substantially different domain and a larger test set, and every rating category in both datasets shows a positive mean Δ . The gradient is not strictly monotonic: on Recipe4U the 5-star tier gains slightly more than the 4-star tier, and we return to this inversion below. Appendix L presents a representative 1-star example.

Two caveats apply to these figures. The Recipe4U low tiers are small (1-star $N=11$, 2-star $N=16$), so those large Δ values carry real uncertainty (2-star $p=0.057$; Table 8, Appendix H) and are unadjusted for multiple comparisons. The HelpSteer2 low tiers are larger ($N=323$, 390), more stable, and show the same monotonic pattern, so we treat the asymmetry as the central observation.

These deltas also carry a calibration signal. The held-out MAE = 0.736 averages over a rating distribution concentrated in the middle and upper tiers (Appendix G), so the regression is best constrained where data is densest and least constrained at the ends of the scale, where retrieval draws on fewer neighbours. The deltas agree: the largest values and widest seed-to-seed spread sit at the extremes, the Recipe4U 1-star tier varying by ± 0.19 across configurations against ± 0.01 –0.03 elsewhere. Part of the movement at 1-star is therefore scorer uncertainty rather than quality change, and we rest the claim on the sign and ordering of the deltas, not the magnitude at any single tier. Table 11 (Appendix N) measures this on a run with known human labels and confirms both the U-shape and its direction: the judge overrates the bottom of the scale and underrates the top. Full probabilistic calibration (ECE, conformal intervals) is left to future work.

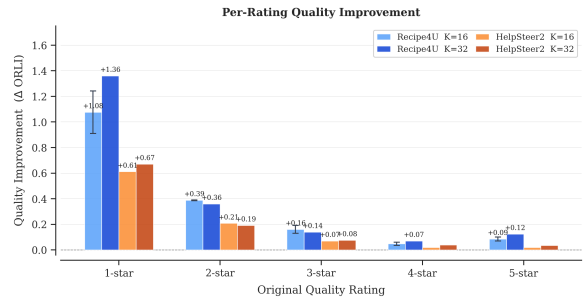


Figure 3: Per-rating Δ ORLI across both datasets and group sizes.

5.2 Comparison with supervised fine-tuning

We compare NORA against supervised fine-tuning (SFT), which isolates pure imitation. SFT gives a modest positive Δ overall, but disaggregation

Dataset	Rating	N	K=16 (s=42)	K=16 (s=123)	K=32 (s=42)	Mean ± SD
Recipe4U	1-star	11	+1.242	+0.911	+1.361	+1.17±0.19
	2-star	16	+0.391	+0.385	+0.360	+0.38±0.01
	3-star	51	+0.129	+0.192	+0.140	+0.15±0.03
	4-star	140	+0.036	+0.060	+0.070	+0.06±0.01
	5-star	164	+0.070	+0.101	+0.124	+0.10±0.02
	Overall	382	+0.112	+0.133	+0.152	+0.13±0.02
HelpSteer2	1-star	323	+0.612	+0.590	+0.671	+0.62±0.03
	2-star	390	+0.208	+0.212	+0.192	+0.20±0.01
	3-star	489	+0.070	+0.075	+0.076	+0.07±0.00
	4-star	1,175	+0.020	+0.023	+0.039	+0.03±0.01
	5-star	1,686	+0.020	+0.022	+0.035	+0.03±0.01
	Overall	4,063	+0.091	+0.092	+0.107	+0.10±0.01

Table 2: Per-rating quality improvement (Δ), largest where original quality is lowest.

exposes the imitation ceiling (Section 3): its 3-star examples degrade by -0.088 on Recipe4U and -0.062 on HelpSteer2. Trained only on high-rated demonstrations, it collapses toward that mode, pulling mid-tier outputs toward the high-rated style without improving substance. NORA avoids this by anchoring early training to gold-response similarity before shifting to quality optimisation, giving positive Δ on every tier. Table 3 gives the full comparison, and Figure 4 (Appendix I) visualises it.

Dataset	Method	Overall Δ	Low (1–2★)	Medium (3★)	High (4–5★)
Recipe4U	SFT Baseline	+0.096	+0.18	−0.088	+0.12
	NORA ($K=16$)	+0.112	+0.74	+0.13	+0.05
	NORA ($K=32$)	+0.152	+0.77	+0.14	+0.10
HelpSteer2	SFT Baseline	+0.081	+0.14	−0.062	+0.09
	NORA ($K=16$)	+0.091	+0.39	+0.07	+0.02
	NORA ($K=32$)	+0.107	+0.41	+0.08	+0.04

Table 3: Comparison with supervised fine-tuning, where bold marks the best value per column. SFT degrades medium-quality responses, whereas NORA achieves positive improvement across all tiers.

5.3 Ablation study

We ran three ablation conditions on both datasets ($K=16$, seed 42), each disabling exactly one mechanism while keeping the rest of the pipeline intact. Table 4 reports the numbers, and Figure 9 (Appendix M) visualises them.

Removing the curriculum ($\alpha=0$, $\beta=1$ throughout) causes the largest drop, with overall Δ falling 46% on Recipe4U and cosine similarity collapsing from 0.591 to 0.482. This shows ORLI, like any learned reward, *can* be over-optimised when unconstrained; the curriculum’s early similarity anchoring is what prevents exploitation, not embedding-based scoring itself. Disabling contrastive regularisation ($\lambda_c=0$) reduces aggregate Δ by only 16–17%, but the low-rated subset drops $\sim 28\%$

on both datasets, consistent with the loss being gated by $r \leq 2$. Replacing rating-aware with uniform retrieval reduces low-rated gains by 38–39%. Detailed per-condition analysis appears in Appendix M.

Dataset	Configuration	Overall Δ	Low (1–2★) Δ	Sim
Recipe4U	Full NORA	+0.112	+0.74	0.591
	w/o Curriculum	+0.061	+0.39	0.482
	w/o Contrastive	+0.094	+0.53	0.603
	w/o Rating-Aware	+0.078	+0.46	0.568
HelpSteer2	Full NORA	+0.091	+0.39	0.548
	w/o Curriculum	+0.049	+0.21	0.461
	w/o Contrastive	+0.076	+0.28	0.559
	w/o Rating-Aware	+0.063	+0.24	0.523

Table 4: Ablation study across both datasets ($K=16$, seed 42). Removing the curriculum causes the largest drop in both overall Δ and cosine similarity.

5.4 Scope of the comparison

Our baselines are gold responses and SFT. We do not compare ORLI against a trained ordinal reward model consuming the same 1–5 labels, or against an LLM-as-a-judge reward under matched conditions, though both are the natural competitors. Our results therefore support the claim that a retrieval-augmented deterministic scorer suffices to drive asymmetric quality gains under GRPO, not the stronger claim that it outperforms either alternative; the comparisons in Tables 1–4 are internal to NORA. Such a comparison also presents design difficulties, since the three scorers do not share a scale and none can serve as the shared metric without reintroducing the circularity we note for ORLI. It would require held-out MAE and per-tier bias against common human labels, V_{gen} measured empirically for the generative judge, and policy quality under each reward assessed by blinded human evaluation. We regard this as the necessary next experiment.

5.5 Human evaluation

ORLI also serves as the training reward, so it may overstate performance. To test whether the asymmetric pattern holds under blinded assessment, we ran a purposive evaluation on nine Recipe4U examples ($K=32$, seed 42): three typical (not extreme) examples per tier (1–2, 3, 4–5 star), each judged by five independent evaluators. Responses were unlabelled and order-randomised, and no evaluator knew of NORA or the hypothesis. Full protocol and per-example votes are in Appendix J.

The asymmetric pattern replicates under human judgement (Table 5). All three low-rated exam-

Tier	N (ex.)	Original (mean $\pm$ SD)	NORA (mean $\pm$ SD)	Human Δ	Win / Tie / Lose (of 15)
1-2 star	3	1.60 $\pm$ 0.42	3.47 $\pm$ 0.51	+1.87	87% / 7% / 7%
3 star	3	3.07 $\pm$ 0.48	3.33 $\pm$ 0.45	+0.27	53% / 33% / 13%
4-5 star	3	4.27 $\pm$ 0.46	4.33 $\pm$ 0.49	+0.07	27% / 47% / 27%
Overall	9	2.98 $\pm$ 1.12	3.71 $\pm$ 0.51	+0.73	56% / 29% / 16%

Table 5: Human evaluation by quality tier: nine Recipe4U examples, five evaluators each.

ples produce a NORA majority, with a human Δ of +1.87 that closely tracks the ORLI delta; evaluators found NORA outputs more structured and actionable. On high-rated examples votes split roughly evenly and the human Δ is a negligible +0.07. Fleiss’ $\kappa = 0.64$ overall (rising to 0.78 on the low-rated tier) indicates substantial agreement (Landis and Koch, 1977); it drops on higher-quality examples, where NORA’s changes are marginal.

6 Conclusion

NORA replaces generative evaluation with deterministic regression over a human-rated embedding manifold, zeroing the generation-induced (epistemic) component of reward uncertainty by construction and leaving approximation bias and the aleatoric uncertainty of the labels. That removal is a reallocation of the error budget rather than a reduction of it, and our per-tier analysis suggests the bias term is now the binding constraint. ORLI remains a learned reward, so NORA relies on a retrieval-guided curriculum to curb over-optimisation; without it, the policy exploits the scorer until coherence collapses. The improvement is asymmetric: NORA gains on every rating tier, largest where quality is lowest, across two datasets, multiple seeds, and two group sizes, confirmed by blinded human evaluation.

Limitations

ORLI is both the training reward and the primary automatic metric, and the blinded human study was purposive on one dataset, so ORLI gains remain a proxy. Per-rating deltas invite a regression-to-the-mean confound; a matched untrained-policy null (Appendix N) shows it is real but does not explain the asymmetry, though the main-text deltas are uncontrolled for it. All main-text runs use a single encoder, leaving encoder sensitivity untested. We report no head-to-head comparison against a trained ordinal reward model or an LLM-as-a-judge, so our results show only that a deterministic retrieval-augmented scorer suffices. NORA also needs dense

human-rated data and captures semantic similarity rather than factual correctness. Appendix O expands on each of these points.

Acknowledgements

We thank the reviewers for their detailed and constructive feedback, which sharpened the claims of this paper, in particular the scope of Proposition 1 and the positioning of ORLI relative to ordinal reward models. We are also grateful to the volunteers who took part in the human evaluation. This work received no external funding.

Ethics statement

This work uses publicly released, permissively licensed datasets, and all conversations in them are anonymised. ORLI inherits biases present in the human annotations and does not remove them. The human evaluation (Section 5.5) involved five adult volunteers, who took part unpaid and gave informed consent. Participation was anonymous: we recorded only ratings and preference labels and collected no personally identifying information, so evaluators cannot be re-identified. The material came from the public Recipe4U test set and contains no sensitive content.

Two risks merit attention from anyone deploying this method. ORLI’s determinism and speed may invite over-trust, so its scores should not be treated as ground truth without independent validation in new domains. Because the embedding space captures semantic similarity rather than factual correctness, it may reward fluent-but-incorrect outputs in safety-critical domains such as medical or legal advice, which warrants stress-testing before deployment.

Reproducibility statement

All experiments were conducted on a single GPU with ≥ 80 GB memory, rented via a cloud compute provider. The base model, encoder, and both datasets are public, and hyperparameters are in Appendix F. All code is publicly released at <https://github.com/Kcbir/nora>.

References

Karin Avnit, Yong En Nah, Akshita Abrol, Aik Beng Ng, and Zhengkui Wang. 2025. [ForumAId: Enhancing student-teacher interaction through GenAI in LMS environments](#). Conference contribution, Singapore Institute of Technology.

- Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, and 32 others. 2022. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*.
- Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 41–48, Montreal, Quebec, Canada. ACM.
- Thomas Coste, Usman Anwar, Robert Kirk, and David Krueger. 2024. Reward model ensembles help mitigate overoptimization. In *International Conference on Learning Representations*.
- DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems*, volume 36.
- Leo Gao, John Schulman, and Jacob Hilton. 2023. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*. PMLR.
- Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910. Association for Computational Linguistics.
- Jieun Han, Haneul Yoo, Junho Myung, Minsun Kim, Tak Yeon Lee, So-Yeon Ahn, and Alice Oh. 2024. [RECIPe4U: Student-ChatGPT interaction dataset in EFL writing education](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13666–13676, Torino, Italy. ELRA and ICCL.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [LoRA: Low-rank adaptation of large language models](#). In *International Conference on Learning Representations*.
- Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2021. [Billion-scale similarity search with GPUs](#). *IEEE Transactions on Big Data*, 7(3):535–547.
- Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2020. [Generalization through memorization: Nearest neighbor language models](#). In *International Conference on Learning Representations*.
- Sungdong Kim, Sanghwan Bae, Jamin Shin, Soyoung Kang, Donghyun Kwak, Kang Min Yoo, and Minjoon Seo. 2023. Aligning large language models through synthetic feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13677–13700. Association for Computational Linguistics.
- J. Richard Landis and Gary G. Koch. 1977. [The measurement of observer agreement for categorical data](#). *Biometrics*, 33(1):159–174.
- Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Lu, Thomas Mesnard, Colton Bishop, Victor Carbone, and Abhinav Rastogi. 2024. RLAI: Scaling reinforcement learning from human feedback with AI feedback. In *International Conference on Machine Learning*. PMLR.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744.
- Arjun Panickssery, Samuel R. Bowman, and Shi Feng. 2024. LLM evaluators recognize and favor their own generations. *arXiv preprint arXiv:2404.13076*.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pages 53728–53741.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-BERT: Sentence embeddings using Siamese BERT-networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.
- Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. 2016. [Prioritized experience replay](#). In *International Conference on Learning Representations*.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.
- Joar Skalse, Nikolaus H. R. Howe, Dmitrii Krasheninnikov, and David Krueger. 2022. Defining and characterizing reward hacking. In *Advances in Neural Information Processing Systems*, volume 35.

Zhilin Wang, Yi Dong, Olivier Delalleau, Jiaqi Zeng, Gerald Shen, Daniel Egert, Jimmy J. Zhang, Makesh Narsimhan Sreedhar, and Oleksii Kuchaiev. 2024. HelpSteer2: Open-source dataset for training top-performing reward models. *arXiv preprint arXiv:2406.08673*.

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

A Future work

The most important open comparison is the head-to-head against a trained ordinal reward model and an LLM-as-a-judge under matched conditions, including an empirical measurement of the V_{gen} term that Proposition 1 removes (Section 5.4). Beyond that, this work was motivated in part by educational assessment, where reproducible and calibrated scoring of student responses matters and a scorer that returns the same mark for the same answer is a requirement rather than a convenience. We will investigate NORA on real LMS discussion posts collected through ForumAid (Avnit et al., 2025), evaluating its reliability and its agreement with human assessment in a deployed setting, a regime that also supplies the dense, pointwise rated dialogue data the method assumes.

B Curriculum reward schedule: design rationale

The curriculum reward (Equation 4) combines a similarity term R_{sim} and a quality term R_{ORLI} through time-dependent weights $\alpha(t)$ and $\beta(t)$. The linear annealing schedule is a design choice motivated by a stability consideration: during early training, the policy must remain close to the data distribution to avoid degenerate outputs, while later training should prioritise quality improvement.

The underlying rationale is as follows. R_{sim} acts as a soft anchor that penalises deviation from gold responses, while R_{ORLI} encourages distributional shift toward higher-quality outputs. We want the anchor to dominate early and the quality signal to dominate late. A linear schedule is the simplest monotonic interpolation that achieves this. We parameterise:

$$\alpha(t) = \alpha_0 - (\alpha_0 - \alpha_T) \cdot \frac{t}{T} \quad (13)$$

$$\beta(t) = \beta_0 + (\beta_T - \beta_0) \cdot \frac{t}{T} \quad (14)$$

where $\alpha_0 = 1.0$, $\alpha_T = 0.2$, $\beta_0 = 0.1$, $\beta_T = 0.8$. These values were chosen so that the quality term dominates at convergence ($\beta_T/\alpha_T = 4$) but the similarity term never vanishes entirely ($\alpha_T > 0$), preventing unbounded drift. Note that $\alpha(t) + \beta(t) \neq 1$ in general; however, GRPO’s group normalisation (Equation 8) renders the absolute scale of R irrelevant, so only the ratio $\beta(t)/\alpha(t)$ matters. This ratio increases monotonically from 0.1 to 4.0 over training.

We do not claim this is the only or optimal schedule. The linear form was chosen for simplicity and interpretability; cosine, exponential, or step-function alternatives might perform comparably or better. The specific endpoint values ($\alpha_T=0.2$, $\beta_T=0.8$) were selected through grid search on a held-out validation split (Appendix F). The ablation in Table 4 confirms that the curriculum mechanism matters substantially, but the precise functional form remains an open design parameter.

C Contrastive loss derivation

The contrastive regularisation term (Equation 7) draws on the principle from SimCSE (Gao et al., 2021) that pushing representations away from known negatives improves the quality of the learned embedding space. In our setting, the “negatives” are not random samples but specifically retrieved low-quality responses from the same rating tier.

For a generated response $\hat{y}$ to a low-rated query ($r \leq 2$), rating-aware retrieval returns k nearest neighbours $Y^- = \{y_1^-, \dots, y_k^-\}$ from tiers $r \in \{1, 2\}$. The contrastive loss is:

$$\mathcal{L}_{\text{contrastive}} = \frac{1}{k} \sum_{j=1}^k \frac{E(\hat{y}) \cdot E(y_j^-)}{\|E(\hat{y})\| \|E(y_j^-)\|} \quad (15)$$

This is identical to Equation 7 in the main text, expanded for clarity. Because cosine similarities in our embedding space are empirically positive, minimising $\mathcal{L}_{\text{contrastive}}$ reduces the similarity between $\hat{y}$ and each y_j^- , pushing the generation away from low-quality response patterns. No negation is needed: gradient descent on a positive similarity directly encourages divergence. The total training objective is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{policy}} + \lambda_c \cdot \mathbb{I}[r \leq 2] \cdot \mathcal{L}_{\text{contrastive}} \quad (16)$$

where $\lambda_c = 0.3$ and $\mathbb{I}[r \leq 2]$ activates the term only for low-rated examples.

D GRPO advantage normalisation

Standard policy gradient methods compute advantages using a learned value function $V_\psi(s)$. GRPO replaces this with group-level normalisation. For a query x , we sample K candidate responses $\{\hat{y}_1, \dots, \hat{y}_K\}$ and compute rewards $\{R_1, \dots, R_K\}$. The group-normalised advantage for candidate i is:

$$A_i = \frac{R_i - \bar{R}}{\sqrt{\frac{1}{K} \sum_{j=1}^K (R_j - \bar{R})^2 + \epsilon}} \quad (17)$$

where $\bar{R} = \frac{1}{K} \sum_{j=1}^K R_j$ and $\epsilon = 10^{-8}$ prevents division by zero.

The policy gradient with PER importance weights becomes:

$$\nabla_{\theta} \mathcal{L} = -w(x) \sum_{j=1}^K A_j \cdot \nabla_{\theta} \log \pi_{\theta}(\hat{y}_j | x) \quad (18)$$

where $w(x)$ is the importance-sampling weight of the query x drawn from the replay buffer, and the same weight multiplies all K candidates in the group.

We follow the standard prioritised-replay formulation of Schaul et al. (2016). A buffer entry i with priority $p_i = (6 - r_i)/5 + 0.2$ is sampled with probability

$$P(i) = \frac{p_i^{\alpha_{\text{PER}}}}{\sum_{i'} p_{i'}^{\alpha_{\text{PER}}}}, \quad (19)$$

so that lower-rated examples are visited more often, and the resulting sampling bias is corrected by

$$w(x_i) = \frac{1}{Z} \left(\frac{1}{|\mathcal{B}| \cdot P(i)} \right)^{\beta_{\text{PER}}}, \quad (20)$$

with $|\mathcal{B}|$ the buffer size and $Z = \max_{i'} w(x_{i'})$ normalising the weights to at most 1 for stability. The exponent β_{PER} is annealed from its initial value to 1 over training, so the correction is partial early, when the gradient estimate is noisy anyway, and complete by convergence. The weights therefore counteract the prioritisation rather than compounding it: prioritisation decides where the policy spends its updates, and w keeps the gradient an unbiased estimate of the uniform-sampling gradient.

E Training algorithm

Algorithm 1 summarises the complete NORA training procedure described in Section 3.

F Hyperparameter settings

Table 6 lists all hyperparameters. Values were selected through grid search on a held-out 10% validation split of the Recipe4U training set and applied without modification to HelpSteer2.

G Dataset statistics

Table 7 provides structural statistics for both evaluation datasets.

Algorithm 1 Complete NORA training procedure, from judge pretraining to policy updates

Require: Dataset $\mathcal{D} = \{(x_i, y_i, r_i)\}$, group size K , total steps T

Ensure: Trained policy π_{θ}

- 1: Pretrain ORLI judge f_{ϕ} on $\mathcal{D}_{\text{train}}$ ▷ Eq. 1
- 2: Build rating-partitioned FAISS indices
- 3: Initialise π_{θ} from Qwen2-0.5B-Instruct + QLoRA
- 4: **for** $t = 1, \dots, T$ **do**
- 5: Compute $\alpha(t), \beta(t)$ ▷ Eqs. 5–6
- 6: **for** each batch $(x, y, r) \sim \text{PER}(\mathcal{D})$ **do**
- 7: Sample K candidates $\{\hat{y}_j\}_{j=1}^K \sim \pi_{\theta}(\cdot | x)$
- 8: Retrieve references Y^{ref} via rating-aware strategy
- 9: **for** $j = 1, \dots, K$ **do**
- 10: $R_j \leftarrow \alpha(t) \cdot \text{sim}(E(\hat{y}_j), E(Y^{\text{ref}}))$
 $+ \beta(t) \cdot f_{\phi}(E(\hat{y}_j))$
- 11: **end for**
- 12: $A_j \leftarrow (R_j - \mu_g) / (\sigma_g + \epsilon)$ ▷ Eq. 8
- 13: **if** $r \leq 2$ **then**
- 14: $\mathcal{L}_c \leftarrow \frac{1}{|Y^-|} \sum \text{sim}(E(\hat{y}), E(y^-))$
- 15: **end if**
- 16: $\mathcal{L} \leftarrow \mathcal{L}_{\text{policy}} + \lambda_c \cdot \mathcal{L}_c$
- 17: Update θ via AdamW with gradient clipping
- 18: **end for**
- 19: **end for**

Component	Parameter	Value
Actor (QLoRA)	Base model	Qwen2-0.5B-Instruct
	LoRA rank r	8
	LoRA α	32
	LoRA dropout	0.05
	Quantisation	4-bit NF4
	Trainable params	~0.5M (0.1%)
ORLI Judge	Hidden dims	768→256→128→1
	Activation	GELU
	Dropout	0.2 / 0.1
	Loss	Huber ($\delta = 1.0$)
	Pretrain epochs	15
	Learning rate	1×10^{-4}
Policy Training	Optimiser	AdamW
	Learning rate	2×10^{-5}
	Gradient clip	1.0
	Epochs	3
	Batches/epoch	80
	Batch size	2
	Group size K	16 or 32
	Temperature range	[0.7, 1.8]
Curriculum	α_0, α_T	1.0, 0.2
	β_0, β_T	0.1, 0.8
	λ_c	0.3
	Retrieval k	3
PER	Priority p_i	$(6 - r_i)/5 + 0.2$
	α_{PER}	0.6
	β_{PER} (start, end)	0.4, 1.0
	ϵ_{PER}	1×10^{-6}
Encoder	Buffer	Stratified by rating
	Model	all-mpnet-base-v2
	Dim	768

Table 6: Complete hyperparameter settings used across all NORA experiments.

Statistic	Recipe4U	HelpSteer2
Total dialogues	1,913	20,324
Train split	1,531 (80%)	16,261 (80%)
Test split	382 (20%)	4,063 (20%)
Rating scale	1–5	0–4 $\rightarrow$ 1–5
Avg. response length (chars)	682	1,474
<i>Rating distribution (full dataset)</i>		
1-star	3%	8%
2-star	4%	10%
3-star	13%	12%
4-star	37%	29%
5-star	43%	41%

Table 7: Dataset statistics for Recipe4U and HelpSteer2.

H Per-tier significance tests

Table 8 reports one-sample t -tests on the per-example quality deltas summarised in Table 2.

Rating	N	Mean Δ	t	p	Cohen’s d
1-star	11	+1.361	4.10	0.002	1.24
2-star	16	+0.360	2.06	0.057	0.52
3-star	51	+0.140	2.07	0.044	0.29
4-star	140	+0.070	3.29	0.001	0.28
5-star	164	+0.124	4.11	<0.001	0.32
Overall	382	+0.152	5.87	<0.001	0.30

Table 8: One-sample t -tests on per-example quality deltas by rating category (Recipe4U, $K=32$, seed 42), testing $H_0: \mu_\Delta = 0$. All categories but 2-star reach $p < 0.05$. The p -values are unadjusted for multiple comparisons; a Bonferroni correction ($\alpha_{\text{adj}} = 0.01$) would render the 2- and 3-star results non-significant. HelpSteer2 is omitted: its larger per-tier samples (Table 2) already give stable estimates.

I Supplementary figures

This appendix collects additional figures that complement the main text. Figure 4 shows the cross-dataset improvement pattern, Figure 5 provides an architectural diagram of the ORLI judge, and Figure 6 illustrates the rating-aware retrieval routing logic.

J Human evaluation protocol

This appendix provides the full protocol details for the human evaluation reported in Section 5.5.

Design rationale. The evaluation is purposive rather than random. With a small pool of volunteer evaluators, distributing effort across many examples reduces per-item judgement quality through fatigue and attention drift. Instead, we select a small number of examples that maximally discriminate between the three quality regimes discussed in the

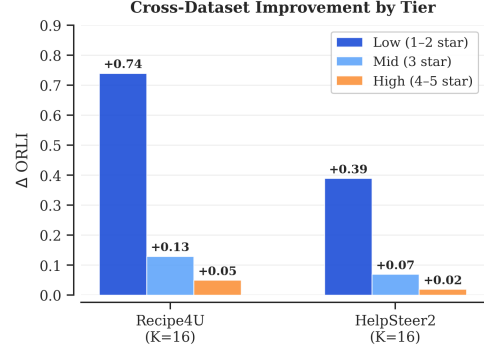


Figure 4: Cross-dataset improvement by quality tier ($K=16$, seed 42). The asymmetric pattern replicates across both datasets: low-rated examples (1–2 star) show the largest gains, medium examples (3 star) show moderate gains, and high-rated examples (4–5 star) change minimally.

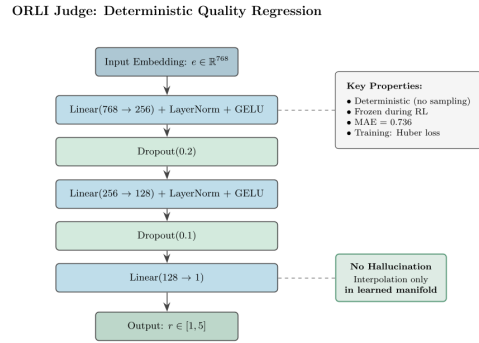


Figure 5: ORLI judge architecture, in which the MLP computes quality scores through deterministic regression: $768 \rightarrow 256 \rightarrow 128 \rightarrow 1$ with GELU activations, LayerNorm, and dropout. Identical inputs always produce identical scores, and the judge attains MAE = 0.736 on held-out validation.

paper and recruit enough evaluators to obtain reliable majority votes with Fleiss’ κ as the agreement statistic. This design directly tests the asymmetry hypothesis rather than attempting population-level representativeness.

Setup. We selected 9 examples from the Recipe4U test set ($K=32$, seed 42): 3 from the 1–2 star tier, 3 from the 3-star tier, and 3 from the 4–5 star tier. Within each tier, examples were chosen to represent typical (not extreme) instances of that rating level, avoiding outliers that might skew judgements in either direction. Five independent evaluators assessed each example. Evaluators received the student query and two responses side-by-side (“Response A” and “Response B”), with the assignment of gold and NORA outputs to labels randomised per example to prevent position bias.

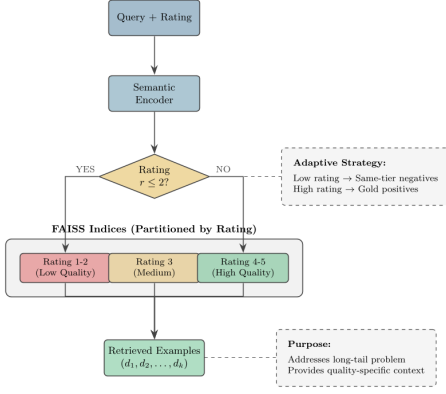


Figure 6: Rating-aware retrieval flowchart, in which queries are routed to different FAISS partitions based on their original rating: low-rated queries ($r \leq 2$) retrieve same-tier negatives for contrastive learning, while higher-rated queries retrieve gold exemplars ($r \geq 4$) for imitation.

They completed two tasks: (1) rate each response on a 1–5 scale for helpfulness, clarity, and pedagogical value; and (2) state which response they preferred (“first better”, “roughly equal”, “second better”). Evaluators were not told which system produced which response, how many systems were involved, or anything about the NORA framework.

Per-example results. Table 9 shows the majority vote and vote distribution for each individual example. The shift from unanimous NORA preference on low-rated examples to divided or original-leaning judgements on high-rated examples mirrors the asymmetric pattern observed in automatic ORLI metrics.

Ex.	Tier	Majority vote	Agreement	Votes (NORA / Tie / Orig)
1	1–2★	NORA	Unanimous (5/5)	5 / 0 / 0
2	1–2★	NORA	Strong (4/5)	4 / 1 / 0
3	1–2★	NORA	Strong (4/5)	4 / 0 / 1
4	3★	NORA	Moderate (3/5)	3 / 2 / 0
5	3★	NORA	Moderate (3/5)	3 / 1 / 1
6	3★	NORA	Marginal (2/5)	2 / 2 / 1
7	4–5★	Divided	Split (2/5 each)	2 / 2 / 1
8	4–5★	Tie	Moderate (3/5)	1 / 3 / 1
9	4–5★	Orig-leaning	Marginal (2/5)	1 / 2 / 2

Table 9: Per-example majority-vote breakdown across nine Recipe4U examples and five evaluators. Agreement decreases from the low-rated tier (where quality differences are salient) to the high-rated tier (where differences are marginal).

K Scaling with group size

Figure 7 shows the effect of increasing group size from $K=16$ to $K=32$. Overall Δ increases from

+0.112 to +0.152 on Recipe4U (+36%) and from +0.091 to +0.107 on HelpSteer2 (+18%), at the cost of doubling wall-clock time from 7 h to 14 h. Gains on low-rated examples are more modest ($\sim 4\%$ on Recipe4U, $\sim 5\%$ on HelpSteer2), suggesting that $K=16$ already captures most of the benefit for the hardest cases.

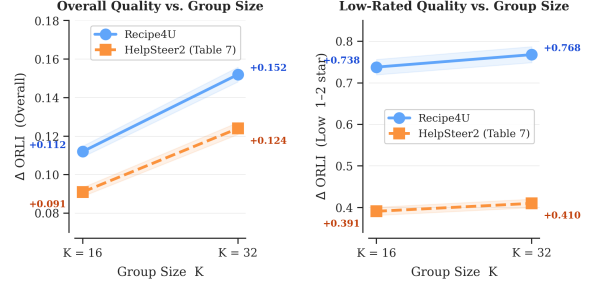


Figure 7: Effect of group size K on quality improvement. Left: overall Δ . Right: low-rated (1–2 star) Δ . Increasing K from 16 to 32 yields diminishing returns on low-rated examples while still improving overall performance.

L Qualitative analysis

To complement the aggregate metrics, we examine a representative 1-star test example from Recipe4U (Figure 8). The original tutor response offers only vague praise (“Great! That’s an excellent summary...”) without actionable guidance. NORA’s output provides structured, pedagogically grounded advice covering paragraph organisation, figure integration, and the distinction between Results and Discussion sections. ORLI rates the generated response at 4/5. While a single example cannot validate the framework, this case illustrates the kind of transformation that drives the aggregate Δ on low-rated examples.

M Ablation analysis details

This appendix provides the per-condition analysis for the ablation study summarised in Section 5.3. Each condition disables exactly one mechanism while keeping the remainder of the pipeline intact ($K=16$, seed 42). Figure 9 visualises the results.

Without curriculum. Setting $\alpha=0$ and $\beta=1$ from the start causes the policy to optimise ORLI scores directly without any similarity anchor. Overall Δ drops by 46% on Recipe4U and by a comparable margin on HelpSteer2. More informative is the collapse in cosine similarity, which falls from 0.591 to 0.482 on Recipe4U and from 0.548 to 0.461

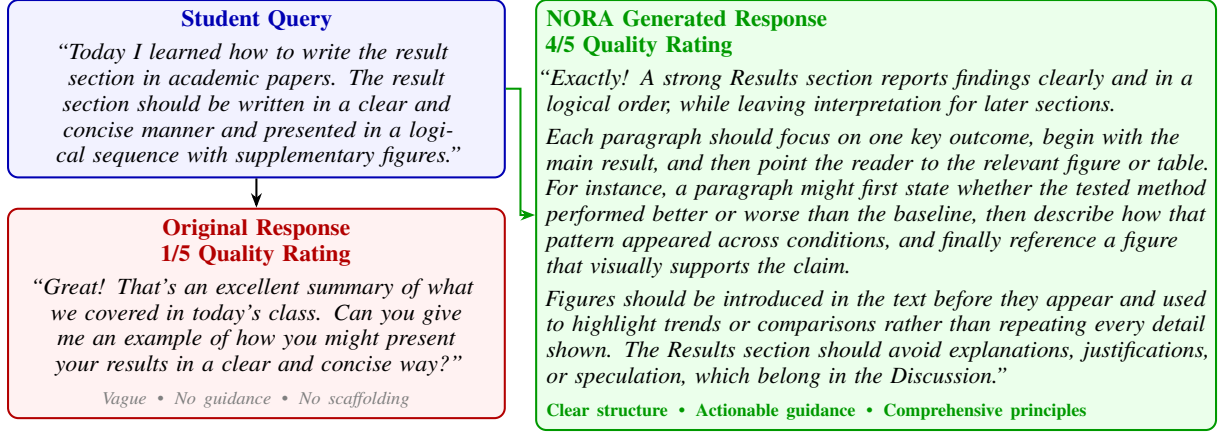


Figure 8: Qualitative comparison on a Recipe4U 1-star example. The original tutor response (left, red) offers only vague praise; NORA’s generated response (right, green) provides structured, actionable guidance drawn from retrieved exemplars.

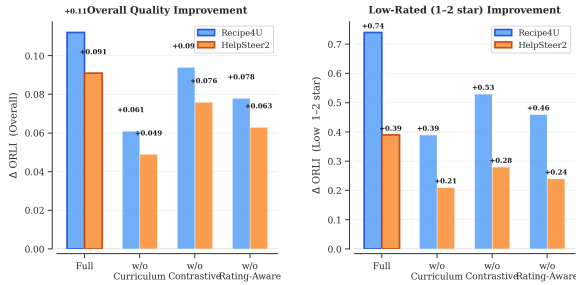


Figure 9: Ablation results visualised. Left: overall Δ ORLI. Right: low-rated (1–2 star) Δ ORLI. The curriculum contributes to both metrics; contrastive regularisation has its effect on the low-rated subset.

on HelpSteer2. Without the early-stage imitation signal, the policy aggressively exploits the ORLI scorer and drifts away from the data distribution, generating outputs that score somewhat higher on the quality metric but lose semantic coherence with the original responses. This is precisely the failure mode the curriculum was designed to prevent, and the ablation confirms that the linear annealing schedule in Equations 5–6 is doing meaningful work.

Without contrastive regularisation. Disabling contrastive regularisation ($\lambda_c=0$) has a smaller aggregate effect, reducing overall Δ by 16–17%. The aggregate figure, however, conceals a targeted effect. On the low-rated subset (1–2 star), Δ drops by roughly 28% on both datasets. The contrastive loss is designed to push generated responses away from known poor examples in embedding space, so its absence has the greatest effect on precisely the tier from which those negatives are retrieved. On high-rated examples the effect is negligible,

which is consistent with the loss being gated by the $\mathbb{I}[r \leq 2]$ indicator.

Without rating-aware retrieval. Replacing rating-aware retrieval with uniform retrieval, where the FAISS query returns neighbours from all tiers regardless of the input’s rating, reduces low-rated gains by 38–39%. This ablation affects both the contrastive loss (which now receives a mix of low- and high-quality negatives) and the similarity reward (which retrieves a less targeted set of gold references). The combined degradation confirms that tier-stratified routing matters most for examples at the extremes of the quality distribution.

N Untrained-policy null control and scaling to a 7B actor

A per-rating delta invites a regression-to-the-mean confound: examples rated 1-star are, in part, examples whose gold response sat at the low end of the noise distribution, so any regenerated response would score higher on average at that tier even with no learning. To separate this from a genuine NORA effect, and simultaneously to check that the asymmetry does not depend on the small 0.5B actor, we reran the full pipeline on HelpSteer2 with a **Qwen2.5-7B-Instruct** actor and a **BGE-large-v1.5** encoder (1024-d), evaluated against an **untrained-policy null**: the same 7B actor with no NORA training and retrieval disabled. Per-tier estimates report bootstrap 95% CIs, Wilcoxon signed-rank tests, and Holm-corrected p -values. The 0.5B Recipe4U results in the main text are unchanged.

Scope of this run. Two details matter for reading this table. First, the evaluation set is capped at 500

examples per rating tier, which is why $N_{\text{test}}=2,190$ here rather than the 4,063 of the main-text HelpSteer2 split: the 4- and 5-star tiers are truncated to 500 while the sparser tiers are unaffected. The cap keeps the tiers closer to balanced and bounds the cost of scoring every candidate, but it means the “Overall” row of Table 10 is not weighted by the natural rating distribution and is not comparable to the overall figures in Table 2. Second, this run is not a minimal edit of the main configuration. In addition to the actor and encoder, it retunes the LoRA adapter (rank 32, $\alpha=64$), the learning rate (1×10^{-5}), the retrieval depth ($k=5$), the judge (20 pretraining epochs, 512 hidden units), the schedule length (500 batches per epoch), and the curriculum endpoints. It is therefore a joint check that the asymmetry survives a scale-up and a matched null, not an ablation isolating any single factor, and its ORLI values live on a different judge’s scale from the main text.

Rating	N	Gold	Null Δ	NORA Δ	NORA–Null
1-star	315	2.896	+0.332	+1.226	+0.894
2-star	388	3.338	+0.112	+0.797	+0.685
3-star	487	3.555	+0.060	+0.621	+0.561
4-star	500	3.751	+0.048	+0.471	+0.423
5-star	500	3.705	+0.053	+0.517	+0.464
Overall	2,190	3.501	+0.104	+0.681	+0.577

Table 10: HelpSteer2 with a 7B actor: NORA against an untrained-policy null. “Null Δ ” is the untrained actor scored against gold; “NORA–Null” removes that regression-to-the-mean component. Bootstrap 95% CIs on the NORA deltas are [1.140, 1.318] (1-star), [0.735, 0.859] (2-star), [0.575, 0.671] (3-star), [0.428, 0.514] (4-star), [0.471, 0.567] (5-star), and [0.554, 0.601] overall. Every NORA delta is significant at Holm-corrected $p < 10^{-50}$.

The asymmetry survives a matched null. Table 10 confirms that the confound is real but does not explain the result. The untrained null reproduces the regression-to-the-mean shape on its own, large at the bottom (+0.33 at 1-star) and negligible at the top (+0.05 at 5-star), which is exactly what a naive per-tier delta would absorb. NORA’s improvement net of this null nonetheless remains largest where quality is lowest (+0.894 at 1-star) and declines toward the top (+0.464 at 5-star), so the asymmetry is not an artefact of conditioning on a noisy baseline. As in the main text the decline is not strictly monotonic, the 5-star tier sitting slightly above the 4-star tier (+0.464 against +0.423). At 7B, NORA also lifts every tier substantially rather

than merely preserving the high ones, which supports the stronger framing that NORA improves every rating tier with the largest gains where quality is lowest.

The main-text results (Section 5) do not include this control, so the per-tier deltas reported there should be read as containing a regression-to-the-mean component of roughly the magnitude in the “Null Δ ” column; the ordering across tiers, which is our actual claim, survives its removal.

The gain is not a length artefact. Embedding-based scorers can reward verbosity, so a policy that simply writes longer answers could produce the pattern above without any quality change. Generation is capped at 192 new tokens (~ 600 characters), and NORA does sit closer to that ceiling than the null: 98.2% of NORA outputs reach it against 85.7% for the untrained policy, giving mean lengths of 598 and 557 characters respectively. To test whether the ceiling explains the result, we restrict to length-matched pairs, keeping only the 1,875 examples where the NORA and null generations differ by at most 25 characters. The gain persists at +0.501 overall against +0.577 unrestricted, and the asymmetry is preserved across tiers (+0.742 at 1-star, +0.587 at 2-star, +0.523 at 3-star, +0.395 at 4-star, +0.400 at 5-star). Length therefore accounts for a modest share of the measured improvement but not for its existence or its shape. Two caveats: the generation cap compresses the length range by construction, so this is a weaker control than it would be with unconstrained decoding, and the main-text 0.5B runs are not length-controlled at all.

Per-tier judge error. Because this run scores 2,190 held-out examples whose human ratings are known, it also lets us measure the judge’s error tier by tier rather than as a single aggregate. Table 11 reports the mean absolute error and the signed bias of the scorer on the gold responses, where the human label is the reference. Two patterns are visible. First, error is not uniform across the scale: it is lowest in the interior (0.49 at 4-star) and rises at both ends (1.90 at 1-star, 1.30 at 5-star), the U-shape anticipated qualitatively in Section 5. Second, and more informative, the error is directional. The judge overrates low-rated responses (+1.90 at 1-star) and underrates high-rated ones (−1.30 at 5-star), compressing predictions toward the interior of the scale. This has a direct bearing on how the per-tier deltas should be read: where the scorer sits well above the human label at 1-star, some of the

measured gain at that tier reflects a floor effect in the scorer rather than a change in response quality, which is a further reason we rest the claim on the ordering of the deltas rather than their magnitude. Note that these figures come from the BGE-large judge used in this 7B run and are therefore not directly comparable to the $\text{MAE} = 0.736$ reported for the all-mpnet judge in the main configuration.

Rating	N	MAE	Signed bias
1-star	315	1.90	+1.90
2-star	388	1.36	+1.34
3-star	487	0.75	+0.55
4-star	500	0.49	−0.25
5-star	500	1.30	−1.30
Overall	2,190	1.09	—

Table 11: Per-tier error of the ORLI judge on gold responses (HelpSteer2, 7B run, BGE-large encoder), with the human rating as reference. Error is lowest in the interior of the scale and higher at both extremes; the signed bias shows the judge overrating low-rated responses and underrating high-rated ones.

O Expanded limitations

This appendix elaborates on each limitation summarised in the Limitations section.

Evaluation circularity. ORLI serves as both the training reward and the primary automatic metric. Although the judge is trained only on training-split labels and frozen before policy optimisation, the policy is optimised to increase ORLI scores, which we then report as our main success measure. The blinded human evaluation (Section 5.5) partially addresses this: five evaluators with no knowledge of NORA confirmed the asymmetric pattern with substantial agreement ($\kappa = 0.64$ overall, 0.78 on the low-rated tier). That study, however, used purposive sampling of nine Recipe4U examples, limiting population-level claims; replication on HelpSteer2 with a larger random sample and a pre-registered protocol would strengthen confidence.

Absent reward-family baselines. The strongest limitation of this study is the set of baselines we do not run. An ordinal reward model trained on the same pointwise labels is the closest competitor to ORLI and shares its determinism, so it isolates exactly the contested variable: whether interpolation via retrieval over a frozen manifold beats interpolation via a network’s learned generalisation. An LLM-as-a-judge reward is the incumbent our mo-

tivation targets, and without it the size of the V_{gen} term we eliminate rests on the literature (Zheng et al., 2023; Panickssery et al., 2024) rather than on our own measurement: we show the term is zero for ORLI but never show how large it is for the alternative, which is what would make zeroing it consequential. Designing the comparison requires care, since the three scorers occupy different scales and none can serve as the shared metric without reintroducing circularity; Section 5.4 sets out the protocol we would use. Until it is run, the practical advantage of the proposed method over an ordinal reward model is unestablished.

Single embedding model. Every main-text experiment uses one encoder, all-mpnet-base-v2, a 110M-parameter sentence transformer. It is unclear how sensitive the results are to this choice: encoders with different objectives, capacities (e.g., e5-large, GTE), or language coverage may organise the space along different quality axes. If an encoder placed high- and low-quality responses closer together, ORLI’s regression surface would change and the asymmetric pattern might not hold. The 7B run in Appendix N does use a different encoder (BGE-large-en-v1.5) and reproduces the asymmetry, but it varies the actor and several training settings at the same time, so it is a joint check rather than an encoder ablation, and its ORLI values are on a different judge’s scale. A controlled sweep holding everything but the encoder fixed would establish robustness bounds; we leave this to future work.

Data density and embedding coverage. NORA’s retrieval design needs sufficiently dense human-rated data. In sparse regimes, where few annotated examples lie near a query, k -nearest-neighbour retrieval degrades and ORLI’s interpolation becomes unreliable; we did not vary the training-set size to characterise this threshold. The embedding space also captures semantic similarity rather than factual correctness: two responses with similar vocabulary but different factual accuracy receive similar embeddings and scores. In code generation, where one wrong operator flips correctness, or mathematical reasoning, where logical validity dominates surface form, ORLI would need pairing with verifiable rewards (e.g., unit tests or proof checkers).