RLEARNER-LLM: BALANCING LOGICAL GROUNDING AND FLUENCY IN LARGE LANGUAGE MODELS VIA HYBRID DIRECT PREFERENCE OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Direct Preference Optimization (DPO) is an efficient alternative to PPO-based RLHF, but on knowledge-intensive tasks its preference signals—from human annotators or LLM judges—exhibit a *verbosity bias* that rewards fluency over logical correctness. This reward-signal blindspot leaves a large logical gap: SFT models reach NLI entailment scores of only 0.05–0.27 despite fluent, confident-sounding text. We propose **RLearner-LLM** and its **Hybrid-DPO**: an automated preference-construction pipeline that fuses a DeBERTa-v3 NLI signal with a verifier LLM score, needing no human annotation while avoiding the "alignment tax" of single-signal optimization. Evaluated across five academic domains (Biology, Medicine, Law) with three base architectures (LLaMA-2-13B, Qwen3-8B, and Gemma 4 E4B-it), RLearner-LLM achieves up to 6.6$\times$ NLI improvement over SFT baselines, with NLI gains in 11 of 15 (architecture, domain) cells and consistent answer-coverage gains on stronger base models. On the 4.5B-effective Gemma 4 E4B-it, Hybrid-DPO lifts NLI in four of five domains (up to +143%; the sole exception, Sydney Biology, already has a high SFT NLI), with faster inference across all five. The gains survive a *strict-discriminator* question-quality filter (keeping only questions whose author key matches the modal student answer): answer-coverage improves in every architecture–corpus cell and NLI in four of six. An independent held-out NLI scorer never used in training reproduces the SFT$\rightarrow$DPO improvement direction in 12 of 15 main cells. A length-controlled follow-up then targets reasoning quality directly: using a pairwise reasoning-quality reward with a length-normalised objective, a compact model's explanations are preferred over its SFT baseline by three independent judges, including the frontier gpt-5-mini (70–82%), at matched output length. Our Qwen3-8B RLearner-LLM wins 95% of blind pairwise comparisons against its own SFT baseline. On the same judge, GPT-4o-mini's longer outputs in turn win 95% against our concise output, and that judge also awards a verbose SFT baseline a 69% win over our DPO-aligned model. These two reversals reproduce the verbosity bias on a frontier comparator, and argue for logic-aware automatic metrics (NLI, ACR) over LLM-as-a-judge on knowledge-intensive generation.[1]

1 INTRODUCTION

The generation of high-quality educational explanations represents a fundamental challenge in natural language processing. Unlike general-purpose conversational tasks, educational rationales must simultaneously satisfy two competing requirements: they must be *logically grounded*, deriving the correct answer through sound deductive steps, and *linguistically fluent*, maintaining natural, readable prose appropriate for a learner audience.

Supervised Fine-Tuning (SFT) on teacher-generated demonstrations has become the standard approach to instilling this capability into LLMs. However, SFT models exhibit a consistent and critical failure mode: while they produce fluent, confident-sounding text, they rarely construct the deductive

[1]Code, training scripts, and per-domain evaluation artefacts (anonymised for review): `https://anonymous.4open.science/r/Explanation-Generation-468D/`

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.

chain that justifies *why* the correct answer follows from the question context. Empirically, SFT models in our study achieve NLI entailment scores of only 0.05–0.27, indicating that the generated explanation rarely logically implies the ground-truth answer, even when the model "knows" the fact.

Reinforcement Learning from Human Feedback (RLHF) and its efficient derivative, Direct Preference Optimization (DPO) (Rafailov et al., 2023), have sought to move beyond SFT by aligning models to human preferences. Yet these approaches contain a structural flaw when applied to factually demanding tasks: **the preference signal does not measure logical correctness**. Human annotators, and LLM judges, exhibit a well-documented *verbosity bias*, systematically preferring longer, rhetorically polished responses over concise, logically precise ones (Zheng et al., 2023). As we demonstrate empirically, GPT-4o-mini awards a 69% win rate to a verbose SFT baseline over a more logically precise but concise RL-aligned model. DPO trained on such signals therefore reinforces stylistic fluency at the expense of factual grounding, leaving the core alignment gap unaddressed.

A natural remedy is to replace human preference labels with an automated logical metric, specifically, NLI entailment probability. However, optimizing DPO solely against NLI entailment incurs what we term the **alignment tax**: the model learns to produce short, repetitive snippets that satisfy the NLI scorer but collapse in linguistic coherence. This trade-off between logical soundness and readability represents a fundamental tension that single-dimensional reward signals cannot resolve.

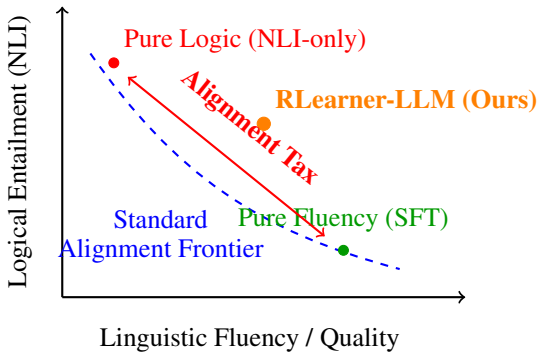


Figure 1: Conceptual illustration of the **Alignment Tax**. Standard optimization forces a trade-off between logical grounding and linguistic fluency. RLearner-LLM pushes the Pareto frontier toward the upper-right quadrant through a dual-signal hybrid reward.

In this paper, we propose **RLearner-LLM**, a framework that resolves this tension by addressing the root cause directly: the reward signal used to construct preference pairs. Our key insight is that logical grounding and linguistic fluency are *complementary objectives*, not competing ones, and that an automated dual-signal reward combining NLI entailment (for logical soundness) with a verifier score (for pedagogical quality) can construct high-fidelity preference pairs without human annotation, while simultaneously avoiding the degenerate solutions that plague single-signal optimization.

Our contributions are as follows:

1. **Problem diagnosis**: We identify that the root cause of logical misalignment in DPO-trained educational models is not the optimization algorithm itself, but the *reward signal blindspot*: the inability of standard preference signals to measure logical correctness. We provide empirical evidence of verbosity bias in LLM-as-a-judge evaluation that perpetuates this blindspot.
2. **Hybrid-DPO framework**: We introduce **RLearner-LLM**, which automatically constructs preference pairs $\mathcal{D}_{pref}$ using a composite score $H(E_i) = 0.5 \cdot S_{\text{NLI}}(E_i) + 0.5 \cdot \tilde{S}_{\text{Verifier}}(E_i)$, fusing a DeBERTa-v3 NLI signal with a min-max-normalised verifier LLM score (Eq. 2). The dual-signal hybrid eliminates the need for human annotation while overcoming the alignment tax.
3. **Cross-architecture validation**: We validate across three base architectures (LLaMA-2-13B, Qwen3-8B, and Gemma 4 E4B-it) and five academic domains (Biology, Medicine, Law), demonstrating up to 6.6× NLI improvement over SFT baselines in 11 of 15 (architecture, domain) cells, with consistent ACR gains on stronger base models; these gains are

paired-Wilcoxon-significant on the lower-baseline trained domains of all three architectures (LLaMA-2-13B on both, Qwen3-8B on Sydney, and Gemma 4 on Cardiff/Medicine; Appendix A). Gemma 4 E4B-it, a 4.5B-effective-parameter dense model, shows Hybrid-DPO NLI gains in four of five domains (+11.9% Auckland Law, +32.0% UK Medicine Y1, +65.6% Cardiff Biology, +143% UK Medicine Y2; not Sydney Biology, already a high SFT NLI), with faster inference across all five and the first single-pass-RL result to surpass the iterative ILearner-LLM (K=5) baseline on Auckland Law NLI. Our Qwen3-8B model further wins 95% of blind pairwise comparisons against its own SFT baseline; on the same judge it loses 95% to GPT-4o-mini's longer outputs, replicating the verbosity bias on a frontier comparator and motivating logic-aware automatic metrics over LLM-as-a-judge for knowledge-intensive generation.

4. **Robustness to question-quality filtering**: Using the full PeerWise answer-submission logs, we show the gains survive a *strict-discriminator* filter that keeps only questions whose author key matches the *modal student answer*: across three architectures and two corpora, answer-coverage improves in all six (architecture, corpus) cells and NLI in four of six (not Qwen3-8B, whose gains shift to coverage), reproducing the per-architecture pattern of the main results (Appendix G).
5. **Independent validation and a step toward reasoning**: An out-of-training held-out NLI scorer (DeBERTa-v3-large) reproduces the improvement direction in 12 of 15 main cells, evidence against reward-hacking of the training scorer. Going beyond entailment, we build a pairwise reasoning-quality reward and, using a length-normalised objective, obtain a compact model whose explanations three independent judges, including the frontier gpt-5-mini, prefer over the SFT baseline (70–82% of non-tied pairs) *at matched output length*, indicating the gain is genuine reasoning quality rather than verbosity (§6.3, Appendix K).

2 Related Work

2.1 Alignment and Preference Optimization

Optimizing Large Language Models (LLMs) to align with human intent has predominantly relied on Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022), utilizing Proximal Policy Optimization (PPO) (Schulman et al., 2017). Recently, Direct Preference Optimization (DPO) (Rafailov et al., 2023) emerged as a stable alternative that mathematically bypasses the need for an explicit reward model. Building on this, recent advancements like RRHF (Yuan et al., 2023), ORPO (Hong et al., 2024), and SimPO (Meng et al., 2024) have further simplified alignment via reference-free objectives. Recent work adds adaptive batch-wise scheduling (Huang et al., 2025) and constrained-alignment methods (Zhang et al., 2025), addressing algorithmic biases in RLHF (Ji et al., 2024; Ma et al., 2024). A parallel line *unifies* SFT and RL into a single objective (Liu et al., 2025; Wang et al., 2024); our contribution is orthogonal and composable with it, keeping the standard SFT→DPO pipeline but repairing the *reward signal* (a dual NLI–verifier reward) and, for reasoning, the *objective* (length-normalised SimPO). Like RLAIF (Lee et al., 2023) our construction needs no human labels; entailment-as-reward has precedent in summarization faithfulness (Pasunuru & Bansal, 2018; Roit et al., 2023), which we adapt to DPO preference construction with a fluency co-signal.

2.2 Hallucination Mitigation and Reasoning

In knowledge-intensive tasks, such as educational explanation generation, factual correctness is paramount. Prior work has explored grounding LLMs using retrieval-augmented generation (RAG) (Lewis et al., 2020) or process-based feedback (Uesato et al., 2022). Modern techniques include mechanistic hallucination detection (Rateike et al., 2023), Multi-Model Contrastive Decoding (Zhu et al., 2025), verifying "decision pivots" in Chain-of-Thought traces (Cho et al., 2025), and cognitive tools (Ebouky et al., 2025) (whether emergent abilities are a metric-driven "mirage" remains debated (Schaeffer et al., 2023)). However, such logical optimizations often degrade fluency, the "alignment tax", which our work addresses via a hybrid verifier-consistent reward.

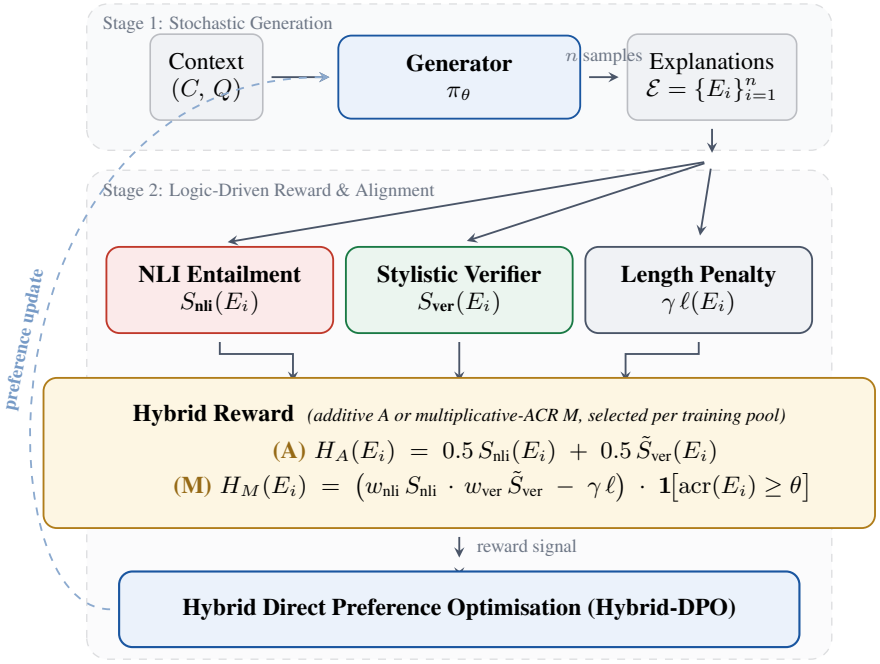


Figure 2: The **RLearner-LLM** framework. Stage 1 samples n candidates from π_θ; Stage 2 scores each with the dual-signal **Hybrid Reward**, either the additive H_A or the multiplicative-ACR H_M (§3.1; selector: pool $\geq \sim$150 pairs and a single/aligned domain $\rightarrow H_M$, else H_A). NLI and verifier feed both variants; the length penalty and ACR gate $\mathbf{1}[\text{acr} \geq \theta]$ enter H_M only. Preference pairs then update π_θ via Hybrid-DPO. $\tilde{S}_{\text{ver}}{=}(S_{\text{ver}}{-}1)/4$ is the $[0,1]$-normalised verifier score.

2.3 LLM-as-a-Judge and Evaluation Bias

The use of powerful instruction-tuned models like GPT-4 as automated evaluators (Zheng et al., 2023) is widespread but prone to significant biases, including verbosity bias (Wang et al., 2023) and self-preference bias (Panickssery et al., 2024). The verbosity-bias confound has been formally analysed in DPO itself: (Park et al., 2024) show that standard DPO and its β-tuned variants (Wu et al., 2024) systematically conflate explanation length with quality, and propose length-aware regularisation as a partial remedy. Our work targets the *upstream* reward-construction stage rather than the downstream loss: by replacing length-correlated stylistic preferences with an NLI-anchored dual-signal reward (Section 3.1) we cut the bias at its origin, which the head-to-head Table 17 confirms persists on frontier comparators despite length controls applied at the output stage. Comprehensive assessments like DecodingTrust (Wang et al., 2023) further highlight the vulnerability of current models to stylistic manipulation; we maintain safety and knowledge-retention standards (Zhang et al., 2025) throughout.

3 Methodology

3.1 First-Principles Derivation of the Hybrid Reward

The RLHF Objective and Its Core Tension. The canonical RLHF objective (Ouyang et al., 2022) is:

$$\max_{\pi_\theta} \mathbb{E}_{y\sim\pi_\theta(y|x)}\big[r(x,y)\big] - \beta\,\text{KL}\big(\pi_\theta \,\|\, \pi_{\text{ref}}\big) \tag{1}$$

where π_{ref} is the supervised fine-tuned reference policy and β controls how much the learned policy is allowed to deviate from it. DPO (Rafailov et al., 2023) derives the closed-form optimal policy $\pi^*(y|x) \propto \pi_{\text{ref}}(y|x) \cdot \exp(r(x,y)/\beta)$ and reformulates the objective as a direct binary cross-entropy loss on preference pairs, eliminating the need for an explicit reward model.

The **alignment tax** arises from a *gradient conflict*: the KL term in Eq. 1 regularises π_θ toward an MLE-trained π_{ref} that prefers verbose, surface-coherent text, while a logic-focused reward (e.g. pure NLI entailment) rewards short dense deductive chains off that fluency manifold. Single-signal optimisation under this conflict either collapses into degenerate snippets that satisfy NLI at the cost of coherence, or capitulates to the KL prior and produces fluent-but-logically-weak text, both manifest as alignment tax on the frontier in Fig. 1.

The Dual-Signal Hybrid Reward. The load-bearing design choice is to score candidates with both an NLI entailment score $S_{\text{NLI}}(E) \in [0,1]$ and a min-max-normalised verifier score $\tilde{S}_{\text{ver}}(E)=(S_{\text{ver}}(E)-1)/4 \in [0,1]$. Used alone each signal degenerates (pure NLI yields answer-repeating snippets; pure verifier reintroduces verbosity bias (Wang et al., 2023; Park et al., 2024)). We instantiate the Hybrid reward at two operating points:

$$H_A(E) = 0.5\, S_{\text{NLI}} + 0.5\, \tilde{S}_{\text{ver}}, \tag{2a}$$

$$H_M(E) = \big(w_{\text{nli}}\, S_{\text{NLI}} \cdot w_{\text{ver}}\, \tilde{S}_{\text{ver}} - \gamma\, \ell_{\text{norm}}\big) \cdot \mathbf{1}[\text{ACR} \geq \theta]. \tag{2b}$$

The additive H_A (Eq. 2a) admits chosen pairs when *either* signal is sufficiently better than the rejected counterpart (high recall). The multiplicative-ACR H_M (Eq. 2b) requires *both* signals to favour the chosen candidate via the product, hard-gates on $\text{ACR}(E) \geq \theta$ to drop answer-evasive candidates, and subtracts a length penalty $\gamma\, \ell_{\text{norm}}(E)$ to suppress the length confounder (high precision). H_A contains no length term: in cross-domain pools the heterogeneous length distribution destabilises a single γ, and H_A is preferred precisely there.

Selector rule. H_M retains $\sim$2–4$\times$ fewer preference pairs than H_A on the same candidate pool (the hard gate drops $\sim 40\%$ of candidates, and the product collapses many surviving pairs below the score-gap threshold). We default to H_M and fall back to H_A when the projected H_M pool drops below ~ 150 pairs or the training corpus merges domains with disparate answer-keyword distributions. This rule is post-hoc rationalisation of our experimental progression; the head-to-head Table 16 shows H_A and H_M within $\sim$2 pp NLI on average across 11 matched cells (mean $H_M - H_A = +1.5$ pp), H_M winning $7/11$. The dual-signal H_M beats both single-signal ablations on Qwen3-8B Cardiff (SFT 0.1959) — NLI-only (0.1419) and verifier-only (0.1368) — while H_A is within noise of NLI-only there; the dual-signal hypothesis, not the algebraic form, is the load-bearing claim.

3.2 Preference Pair Construction

Given n candidate explanations $\{E_i\}_{i=1}^n$ sampled from π_θ for (C, Q), we score each with $H(E_i)$ and retain pairs (E_i, E_j) with $H(E_i) > H(E_j) + \Delta$ as (chosen, rejected), where $\Delta \geq 0.05$. Under H_M, the ACR gate is a hard pre-filter: only pairs whose chosen member passes $\text{ACR} \geq \theta$ are eligible, ruling out the rank-inversion edge case where a long but valid candidate could receive $H_M(E) < 0$ from the length penalty (we verified empirically that this corner case produces no inversions in our 5-domain pool).

3.3 Hybrid-DPO Training Objective

The Hybrid-DPO objective follows the standard DPO formulation (Rafailov et al., 2023) but is trained on $\mathcal{D}_{\text{pref}}$ constructed with the Hybrid reward (Eq. 2):

$$\mathcal{L}_{\text{DPO}}(\pi_\theta) = -\mathbb{E}_{(x, y^+, y^-) \sim \mathcal{D}_{\text{pref}}} \left[\log \sigma \left(\beta \left(\log \frac{\pi_\theta(y^+|x)}{\pi_{\text{ref}}(y^+|x)} - \log \frac{\pi_\theta(y^-|x)}{\pi_{\text{ref}}(y^-|x)} \right) \right) \right] \tag{3}$$

Because every chosen example y^+ in $\mathcal{D}_{\text{pref}}$ outscores its rejected counterpart on *both* dual-signal axes, NLI entailment and verifier fluency, regardless of which Hybrid variant (H_A or H_M) constructed the pair, the gradient of $\mathcal{L}_{\text{DPO}}$ consistently points toward higher logical entailment without sacrificing pedagogical quality. This is the mechanism by which the Hybrid reward breaks the fluency attractor that causes alignment tax in single-signal optimisation.

Implementation notes. Two base models in this study require non-default LoRA target selection. Qwen3-8B applies RMSNorm to q, k after the projection reshape, so LoRA adapters on `q_proj`/`k_proj` would corrupt that path; we restrict its targets to `v_proj`,

o_proj, and the three MLP projections. Gemma 4 E4B-it ships as a multimodal Gemma4ForConditionalGeneration checkpoint whose text-tower weights live under model.language_model.*; we therefore load that class explicitly and use a full-path LoRA target regex that matches exactly the 258 plain nn.Linear attention/MLP projections in the text tower (full pattern in Appendix C). At inference, the DPO adapter must be composed on top of the SFT-merged base (base → apply SFT → merge → apply DPO → merge), since the DPO delta was trained relative to that reference point.

4 Experimental Setup

We conduct experiments on three base architectures spanning two parameter regimes and three pre-training recipes: **LLaMA-2-13B**, **Qwen3-8B**, and **Gemma 4 E4B-it** (4.5B effective parameters via Per-Layer Embeddings). All training is performed on a cluster of NVIDIA A100 GPUs using the TRL and PEFT libraries. The same iterative-refinement **ILearner-LLM (K=5)** baseline of (Bao et al., 2025) is reproduced from a LLaMA-2-13B SFT checkpoint with $K=5$ refinement passes per question, identical to the original publication; we use this as a strictly stronger inference-time baseline ($5\times$ the inference compute of our single-pass models).

4.1 Datasets and Preprocessing

The evaluation spans five academic domains: Cardiff Biology, Sydney Biology, Auckland Law, and UK Medicine (Years 1 and 2). For each domain, we utilize a 100-question held-out test set. Input contexts are formatted as multi-choice questions with associated correct options. The SFT corpus consists of 13,211 question–explanation pairs authored by **undergraduate students** on the PeerWise peer-learning platform. These explanations reflect authentic learner reasoning rather than expert-curated answer keys, which introduces a known ceiling on ground-truth quality (incomplete justifications, occasional misconceptions, and stylistic variance); we quantify the downstream effect of this ceiling through the SFT failure-mode analysis in Table 8 and discuss its implications for future expert-verified corpora in Section 6.4.

4.2 Training Configurations

SFT: 3 epochs on 13,211 examples, LoRA $r=16$, $\alpha=32$, per-device batch size 2, gradient accumulation 8, bf16. **Preference data**: 500 prompts $\times$ $n=3$ candidates scored with $H(E)$ (Eq. 2). H_M uses $w_{\text{nli}}=0.7$, $w_{\text{ver}}=0.3$, $\gamma=0.002\times(\text{word_count}/100)$, $\theta=0.5$ (ACR-failing candidates get zero reward); H_A uses $0.5/0.5$ weights on S_{NLI} and $\tilde{S}_{\text{ver}}=(S_{\text{ver}}-1)/4$ with no gate or length term. Pairs with score gap $\Delta\geq 0.05$ are retained. **DPO**: 5 epochs, lr 5×10^{-5}, batch size 1, grad accum 8, bf16, $\beta_{\text{DPO}}=0.1$. **Variant assignment** (selector rule, §3.1): LLaMA-2-13B uses H_A on the cross-domain pool (239 pairs); Qwen3-8B uses H_M at $N=5$ on Cardiff/Sydney (143 pairs) and H_A cross-domain elsewhere (632 pairs); Gemma 4 E4B-it uses H_M on the 5-domain merged pool (164 pairs). Full assignment in Table 9 (Appendix E).

4.3 Evaluation Tier

We evaluate on four metrics. **Textual Overlap**: BLEU and BERTScore vs. the student reference rationale. **Answer Anchoring**: BERTScore(Ans) and Answer Coverage Rate (ACR) vs. the correct-option text. **Logical Entailment (NLI)**: $S_{\text{NLI}}=P(\text{entail} \mid E, A)\in[0,1]$ from cross-encoder/nli-deberta-v3-small (premise E, hypothesis A, joint max length 512, no prompt wrapping). **Verifier score (Ver)**: raw mean of the Alpaca-7B Likert scorer of (Bao et al., 2025), reported as a diagnostic that the same automatic judge that produced the training signal does not penalise our outputs; the verifier is *not* used as a primary evaluation comparator.

5 Analysis and Results

We compare **RLearner-LLM** against four baselines: (i) each architecture's *SFT* checkpoint; (ii) our verifier-only *DPO v1*/*DPO v2* ablations (no NLI signal) that surface the alignment tax; (iii) the iterative *ILearner-LLM (K=5)* of (Bao et al., 2025) (§4); and (iv) a proprietary *GPT-4o-mini* blind

pairwise judge. RLearner-LLM improves NLI over SFT in 11 of 15 (architecture, domain) cells (the four non-improvements: Qwen3 on Cardiff/Auckland Law/Med Y1, Gemma 4 on Sydney), surpasses ILearner-LLM on four of five domains, and wins 95% of pairwise comparisons against its own SFT baseline while losing 95% to GPT-4o-mini's longer outputs (§5.3, Table 17). Table 1 gives the headline; §§5.1–5.3 unpack each comparison.

Domain	LLaMA-2-13B SFT → **RLearner**	Qwen3-8B SFT → **RLearner**	Gemma 4-E4B-it SFT → **RLearner**	ILearner-LLM (K=5) (iterative ref.)
Cardiff Biology	0.0555 → **0.3209** (5.8×)	0.1959 → 0.1820	0.2117 → **0.3505** (+66%)	0.0864
Sydney Biology	0.0537 → **0.3562** (6.6×)	0.1737 → **0.2284** (+31%)	0.2469 → 0.2309	0.0837
Auckland Law	0.2702 → **0.3229** (+19%)	0.3191 → 0.2303	0.3911 → **0.4377*** (+12%)	0.3996
UK Medicine Y1	0.0860 → **0.4251** (4.9×)	0.2457 → 0.2104	0.2962 → **0.3910** (+32%)	0.1219
UK Medicine Y2	0.2319 → **0.3885** (+68%)	0.1632 → **0.2009** (+23%)	0.1604 → **0.3892** (+143%)	0.0668

Table 1: NLI entailment improvement summary (SFT → RLearner-LLM; bold marks an improvement, parentheses give the relative gain). RLearner-LLM improves NLI in 11 of 15 cells; the four non-improvements (Qwen3 on Cardiff/Auckland Law/Medicine Y1, Gemma 4 on Sydney) are domains whose SFT NLI is already high and where gains shift onto answer-coverage (§6.2; per-domain breakdown in Appendix A). The asterisk marks the first single-pass-RL result to surpass ILearner-LLM (K=5) (Gemma 4, Auckland Law). Paired-Wilcoxon significance for the Gemma-4 flagship and for the LLaMA-2-13B and Qwen3-8B per-domain gains is reported in Appendix A.

5.1 Comparison across Scientific Domains

Table 2 gives the full per-architecture breakdown on Cardiff Biology (other Biology and Medicine domains in Appendix A). On Cardiff, LLaMA-2 RLearner lifts NLI 5.8× over SFT, Qwen3 trades a small NLI delta for a substantial ACR gain (+8.5 pp), and Gemma 4 E4B-it, despite 4.5B effective parameters, achieves the highest NLI among RLearner variants (0.3505, +65.6% over SFT) at the fastest inference (4.76s). On Sydney, LLaMA-2 reaches a 6.6× lift, Qwen3 gains +31.5% NLI / +23.5 pp ACR, and Gemma 4 SFT already sits at a high SFT NLI (0.2469, upper third of our cells); we connect the flat Gemma 4 result to entailment-headroom in §6.2.

5.2 Cross-Domain Performance in Medicine and Law

Zero-shot transfer of the same Hybrid-DPO policy to Auckland Law and UK Medicine, neither contributed questions to the preference-data construction, is summarised in Table 1 (per-architecture detail in Appendix A). **Auckland Law** is the hardest single-pass-RL domain: LLaMA-2 RLearner reaches 0.3229 (+19.5% over SFT) but falls short of ILearner-LLM (K=5) at 0.3996, while Qwen3 delivers a +28.9 pp ACR jump (0.5175→0.8070). Most notably, **RLearner-LLM (Gemma 4 E4B-it)** is the first single-pass-RL result to surpass the ILearner-LLM (K=5) benchmark on Auckland Law (0.4377 vs 0.3996). On **UK Medicine**, all three architectures improve NLI; LLaMA-2 peaks at 0.4251 (Y1) and 0.3885 (Y2), and Gemma 4 matches the LLaMA-2 peak on Y2 (0.3892) with ∼3× fewer effective parameters.

5.3 Pairwise Evaluation and the Persistent Verbosity Bias

Three pairwise comparisons (Table 17) probe LLM-as-a-judge behaviour. Row 1 replicates verbosity bias (Wang et al., 2023; Zheng et al., 2023; Park et al., 2024): GPT-4o-mini awards 69% to the verbose *LLaMA-2 SFT* over the more concise, logically stronger *DPO v2*. Row 2 reverses this pattern when both candidates come from the same family: **RLearner-LLM (Qwen3)** wins 95% vs. its own Qwen3-8B SFT. Row 3 is the cautionary finding: against the proprietary *GPT-4o-mini* comparator, our concise RLearner-LLM *loses* 95% under the same GPT-4o-mini judge. We read this not as a quality verdict but as a continuation of the verbosity bias on a frontier comparator, since GPT-4o-mini's longer hedged outputs satisfy the same length heuristic that drives Row 1.

The juxtaposition of rows 2 and 3 is the headline takeaway: the alignment tax is a property of the *reward signal* used during preference-data construction, not of DPO itself. The Hybrid reward lets an 8B local model beat its own SFT on the same judge, while the same judge reverts to verbosity

bias against a frontier comparator, arguing for logic-aware automatic metrics (NLI, ACR) over LLM-as-a-judge as ground truth on logical quality.

5.4 Reward-Variant Ablation: H_A vs. H_M

H_M wins $7/11$ NLI head-to-heads (mean $+1.5$ pp, Table 16); H_A wins the remaining 4 (three on Sydney). Both variants beat SFT wherever the dual signal is supplied, supporting the view that the dual-signal hypothesis, not the algebraic form, is the load-bearing claim: H_A vs. H_M is a hyperparameter chosen by pair-pool feasibility (§3.1).

5.5 SFT Failure Mode Analysis

A systematic error analysis on $5{,}100$ SFT explanations (Appendix D) finds standard SFT fluent-but-vacuous in 48–85% of cases ($\text{NLI}<0.05$), hallucinating citation-style URLs in 44–69%, and failing to anchor the answer in 12–37% ($\text{ACR}=0$). Hybrid-DPO targets each by construction: the NLI signal suppresses vacuous generations, and under H_M the ACR gate removes answer-evasive candidates while the length penalty discourages verbose hallucinations. A case study (Appendix B) shows the same in microcosm: SFT lists facts without a deductive chain ($\text{NLI}=0.001$) whereas RLearner-LLM produces a structured proof ($\text{NLI}=0.943$).

6 Discussion

6.1 Why the Hybrid Signal Works

NLI verifies that E implies the correct option; the verifier captures clarity invisible to NLI. Each alone degenerates (answer-repeating snippets or verbosity bias), so the joint reward (Eq. 2) filters both extremes and steers the policy to the alignment frontier's upper-right (Fig. 1). The chosen/rejected NLI gap of 17–$18\times$ alongside indistinguishable verifier scores replicates across all three architectures, confirming NLI as the discriminating signal and the verifier as a fluency regulariser. **Robustness to question-quality filtering.** A stricter Tier-B Cardiff filter (`deleted`=0, $\geq 10\%$ 5-star endorsement) shrinks the pool $7\times$ yet *lifts* Hybrid-DPO NLI by $+48\%$ and ACR by $+13.3$ pp (Appendix F). Stacking a Tier-C *strict-discriminator* filter (drop questions whose author key differs from the modal student answer) preserves the pattern: across all six architecture–corpus cells every cell improves ACR and four improve NLI, with the largest lifts on LLaMA-2-13B ($+219\%$/$+204\%$), while Qwen3-8B stays ACR-up/NLI-flat (Appendix G).

6.2 Architecture Effects and Entailment Headroom

Hybrid-DPO's marginal NLI gain is inversely related to the SFT NLI baseline: across all 15 (architecture, domain) cells the relative NLI gain correlates with the SFT NLI baseline at Pearson $r=-0.81$ ($p=3\times10^{-4}$; Fig. 3), so the benefit is set by *entailment headroom*, not architecture or scale. LLaMA-2-13B (lowest SFT NLI, 0.05–0.27) sees the largest lifts (peak $+0.339$/$4.9\times$); Qwen3-8B (higher SFT NLI) shifts gains onto ACR ($+8$ to $+29$ pp); and Gemma 4 E4B-it (4.5B effective) improves NLI on 4 of 5 domains, matching the LLaMA-2 peak with $\sim3\times$ fewer parameters and being the first single-pass result to surpass ILearner-LLM (K=5) on Auckland Law. The lone non-improvement, Sydney Biology, has an already-high SFT NLI (0.2469, upper third of our cells). Since H_M's ACR gate is a floor not a ceiling, a stronger ACR weight is a natural way to recover Qwen3's ACR profile. **Recency control (newer model).** Repeating the Cardiff pipeline on Llama-3-8B (newer and smaller than LLaMA-2-13B) lifts an already-high SFT baseline (NLI 0.305, ACR 0.796) to NLI 0.551 ($+80.7\%$) and ACR 0.876 ($+8.0$ pp), so the SFT$\rightarrow$DPO improvement holds on a modern architecture, not only the older LLaMA-2-13B (NLI-only DPO here; the verifier was unavailable on that machine).

6.3 Directly Optimising Reasoning Quality

Answer entailment (NLI) and coverage (ACR) are proxies: an explanation can entail the correct option by restating it, without giving a valid reason. We therefore test whether the policy can be

trained to produce *better reasoning*, not merely higher entailment, and whether any gain survives a length control, since verbose outputs inflate LLM-judge win rates.

Reasoning-quality reward. From the Cardiff Tier-C training prompts we sample three on-policy explanations per question from the Gemma 4 SFT model and rank them with a *pairwise* reasoning judge (DeepSeek-R1-32B) asked which explanation better justifies *why* the stated answer is correct, explicitly ignoring length and formatting, with order randomised to blunt position bias. This yields 321 preference pairs. To remove a length confound at the source we keep the length-matched subset (112 pairs; chosen 46 vs rejected 44 words). An independent judge from a different family (Llama-3-8B) agrees on 69% of these pairs, indicating a cross-judge preference rather than one model's idiosyncrasy or a length artefact.

Length-preserving optimisation. Plain sequence-level DPO on this reward drifts the policy toward the base model's verbose register; we instead use SimPO, whose length-normalised, reference-free objective removes the incentive to lengthen. The resulting policy stays concise: 68 words on the held-out test versus 66 for SFT (plain DPO 60; reward-filtered self-distillation 44).

Held-out evaluation. On all 191 Cardiff Tier-C test questions, and at matched length, three judges independent of the reward (Qwen2.5-7B, Llama-3-8B, and the frontier gpt-5-mini) prefer the SimPO policy's reasoning to SFT's, by 70%, 72%, and 82% of non-tied pairs ($p \ll 10^{-3}$, sign test); a reward-family reference judge (DeepSeek-R1) agrees at 80%. Plain DPO and the reward-filtered self-distillation variant are weaker or mixed (Appendix K, Table 14). **Seed stability.** Retraining SimPO under three independent random seeds gives a Qwen2.5 non-tied win rate of $76.4 \pm 0.8\%$ (range 75.3–77.1%), so the advantage is a stable property of the method rather than the noise of a single run.

Human study. A blinded study with four independent raters on the 24 length-matched pairs corroborates the automatic judges' direction: raters prefer the SimPO policy in 63% of non-tied judgements (56 vs 33; sign test $p \approx 0.02$; per-item majority 13 vs 5). The margin is smaller than for the LLM judges and inter-rater agreement is modest (49%), reflecting the subjectivity of fine-grained reasoning judgements (Appendix N).

Scope. Two judges are reward-independent and every comparison is length-matched; larger human validation and extension beyond Gemma 4/Cardiff remain future work.

6.4 Limitations

Four limitations apply. (i) LLaMA-2 on Auckland Law (NLI 0.3229) is surpassed by ILearner-LLM (K=5) at 0.3996; Gemma 4 closes the gap (§6.2). (ii) **Circular evaluation**: NLI uses the same `deberta-v3-small` that scored training pairs. An independent held-out scorer (DeBERTa-v3-large) agrees in $12/15$ main cells and $17/21$ overall (all five Gemma-4 domains), evidence against reward-hacking; it does not corroborate some LLaMA-2 gains, so we scope the headline claim to entailment/coverage (Appendix H). (iii) The PeerWise SFT corpus is undergraduate-authored; partial justifications cap overlap metrics, motivating an expert-curated rerun. (iv) **Scope: the benefit tracks entailment headroom** ($r{=}-0.81$, §6.2): where SFT is already strong the gain shrinks or shifts to coverage (Qwen3-8B; Sydney), so Hybrid-DPO most helps weak or learnersourced generators, not ones whose SFT explanations are already answer-faithful.

7 Conclusion

RLearner-LLM closes DPO's reward-signal blindspot with a dual-signal Hybrid reward (NLI + verifier; H_A/H_M by pair-count feasibility). Across 5 domains and 3 architectures it improves NLI in $11/15$ cells (up to $6.6\times$ over SFT) and is the first single-pass-RL method to surpass ILearner-LLM (K=5) on Auckland Law. Going beyond entailment, a length-normalised reasoning reward (SimPO) yields explanations that three independent judges and human raters prefer to SFT *at matched length*. A single quantity organises these results: the NLI gain tracks the SFT baseline at $r{=}-0.81$, so the method both delivers its largest lifts and defines its own scope — it most repairs weak or learnersourced explanation generators, while already-strong baselines shift their gains onto answer coverage. Expert-curated corpora and larger human studies are the natural next steps toward reasoning-faithful educational generation.

Reproducibility Statement

All code, LoRA adapters, preference data, and per-domain evaluation artefacts are released at an anonymised repository (abstract footnote). Exact HuggingFace identifiers, parameter counts, and checkpoints are in Appendix I; SFT/DPO/SimPO hyperparameters, LoRA target modules, and the score-gap filter are in §4.2 and Appendix E; NLI/verifier inputs and the LLM-judge prompts and decoding settings are in Appendix J. All results use the fixed held-out test splits of §4.3; the reasoning-reward construction, length-matching, and multi-judge protocol are detailed in §6.3 and Appendix K.

Ethics Statement

The SFT corpus consists of multiple-choice explanations authored by undergraduate students on the PeerWise peer-learning platform, used for research on educational explanation generation; it contains learner-authored study content rather than personal data. The human evaluation used volunteer, domain-literate raters who judged anonymised model outputs; only optional initials were recorded (reported as R1–R4). RLearner-LLM is intended to *assist* rather than replace educators: generated explanations can still contain factual errors and should be reviewed before classroom use, and we report failure modes (hallucinated citation-style URLs, non-improvement cells, and judge disagreement) transparently throughout.

References

Long Ouyang, Jeff 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 Christiano, Jan Leike, and Ryan Lowe. *Training language models to follow instructions with human feedback*. In Advances in Neural Information Processing Systems (NeurIPS), volume 35, pages 27730–27744, 2022. arXiv:2203.02155.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. *Proximal policy optimization algorithms*. arXiv:1707.06347, 2017.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. *Direct preference optimization: Your language model is secretly a reward model*. In Advances in Neural Information Processing Systems (NeurIPS), volume 36, 2023. arXiv:2305.18290.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. *Retrieval-augmented generation for knowledge-intensive NLP tasks*. In Advances in Neural Information Processing Systems (NeurIPS), volume 33, pages 9459–9474, 2020. arXiv:2005.11401.

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. *Solving math word problems with process- and outcome-based feedback*. arXiv:2211.14275, 2022.

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*. In Advances in Neural Information Processing Systems (NeurIPS), volume 36, 2023. arXiv:2304.05302.

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 (EMNLP), pages 11170–11189, 2024. arXiv:2403.07691.

Yu Meng, Mengzhou Xia, and Danqi Chen. *SimPO: Simple preference optimization with a reference-free reward*. In Advances in Neural Information Processing Systems (NeurIPS), volume 37, 2024. arXiv:2405.14734.

Junkang Wu, Yuexiang Xie, Zhengyi Yang, Jiancan Wu, Jinyang Gao, Bolin Ding, Xiang Wang, and Xiangnan He. *β-DPO: Direct preference optimization with dynamic β*. In Advances in Neural Information Processing Systems (NeurIPS), volume 37, 2024. arXiv:2407.08639.

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. *Judging LLM-as-a-judge with MT-Bench and Chatbot Arena*. In Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track, volume 36, 2023. arXiv:2306.05685.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. *Large language models are not fair evaluators*. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL), pages 9440–9450, 2024. arXiv:2305.17926.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. *LLM evaluators recognize and favor their own generations*. In Advances in Neural Information Processing Systems (NeurIPS), volume 37, 2024. arXiv:2404.13076.

Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, Sang T. Truong, Simran Arora, Mantas Mazeika, Dan Hendrycks, Zinan Lin, Yu Cheng, Sanmi Koyejo, Dawn Song, and Bo Li. *DecodingTrust: A comprehensive assessment of trustworthiness in GPT models*. In Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track (Outstanding Paper), volume 36, 2023. arXiv:2306.11698.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. *Are emergent abilities of large language models a mirage?* In Advances in Neural Information Processing Systems (NeurIPS), volume 36, 2023. arXiv:2304.15004.

Miriam Rateike, Celia Cintas, John Wamburu, Tanya Akumu, and Skyler Speakman. *Weakly supervised detection of hallucinations in LLM activations*. In NeurIPS 2023 Workshop on Socially Responsible Language Modelling Research (SoLaR), 2023. arXiv:2312.02798.

Qiming Bao, Juho Leinonen, Alex Yuxuan Peng, Wanjun Zhong, Gael Gendron, Timothy Pistotti, Alice Huang, Paul Denny, Michael Witbrock, and Jiamou Liu. *Exploring iterative enhancement for improving learnersourced multiple-choice question explanations with large language models*. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI/EAAI-25), pages 28955–28963, 2025. arXiv:2309.10444.

Dongkyu Cho, Amy B. Z. Zhang, Bilel Fehri, Sheng Wang, Rumi Chunara, Hengrui Cai, and Rui Song. *Correct reasoning paths visit shared decision pivots*. In NeurIPS 2025 Workshop on Foundations of Reasoning in Language Models (FoRLM), 2025. arXiv:2509.21549.

Zixuan Huang, Yikun Ban, Lean Fu, Xiaojie Li, Zhongxiang Dai, Jianxin Li, and Deqing Wang. *Adaptive batch-wise sample scheduling for direct preference optimization*. In Advances in Neural Information Processing Systems (NeurIPS), volume 38, 2025. arXiv:2506.17252.

Botong Zhang, Shuo Li, Ignacio Hounie, Osbert Bastani, Dongsheng Ding, and Alejandro Ribeiro. *Alignment of large language models with constrained learning*. In Advances in Neural Information Processing Systems (NeurIPS), volume 38, 2025. arXiv:2505.19387.

Chenyu Zhu, Yefeng Liu, Hao Zhang, Aowen Wang, Yangxue, Guanhua Chen, Longyue Wang, Weihua Luo, and Kaifu Zhang. *Alleviating hallucinations in large language models through multi-model contrastive decoding and dynamic hallucination detection*. In Advances in Neural Information Processing Systems (NeurIPS), volume 38, 2025.

Brown Ebouky, Andrea Bartezzaghi, and Mattia Rigotti. *Eliciting reasoning in language models with cognitive tools*. In Advances in Neural Information Processing Systems (NeurIPS), volume 38, 2025. arXiv:2506.12115.

Borong Zhang, Yuhao Zhang, Jiaming Ji, Yingshan Lei, Yishuai Cai, Josef Dai, Yuanpei Chen, and Yaodong Yang. *SafeVLA: Towards safety alignment of vision-language-action model via constrained learning*. In Advances in Neural Information Processing Systems (NeurIPS), Spotlight, volume 38, 2025. arXiv:2503.03480.

Jiaming Ji, Boyuan Chen, Hantao Lou, Donghai Hong, Borong Zhang, Xuehai Pan, Juntao Dai, Tianyi Qiu, and Yaodong Yang. *Aligner: Efficient alignment by learning to correct*. In Advances in Neural Information Processing Systems (NeurIPS), 2024. arXiv:2402.02416.

Haozhe Ma, Kuankuan Sima, Thanh Vinh Vo, Di Fu, and Tze-Yun Leong. *Reward shaping for reinforcement learning with an assistant reward agent*. In International Conference on Machine Learning (ICML), 2024.

Ryan Park, Rafael Rafailov, Stefano Ermon, and Chelsea Finn. *Disentangling length from quality in direct preference optimization*. In Findings of the Association for Computational Linguistics: ACL 2024, pages 4998–5017, 2024. arXiv:2403.19159.

Mingyang Liu, Gabriele Farina, and Asuman Ozdaglar. *UFT: Unifying supervised and reinforcement fine-tuning*. arXiv:2505.16984, 2025.

Zhichao Wang, Bin Bi, Zixu Zhu, Xiangbo Mao, Jun Wang, Shiyu Wang, Cheng Wang, Dong Nie, and Lingzi Hong. *UFT: Unifying fine-tuning of SFT and RLHF/DPO/UNA through a generalized implicit reward function*. arXiv:2410.21438, 2024.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. *RLAIF: Scaling reinforcement learning from human feedback with AI feedback*. In International Conference on Machine Learning (ICML), 2024. arXiv:2309.00267.

Paul Roit, Johan Ferret, Lior Shani, Roee Aharoni, Geoffrey Cideron, Robert Dadashi, Matthieu Geist, Sertan Girgin, Léonard Hussenot, Orgad Keller, Nikola Momchev, Sabela Ramos, Piotr Stanczyk, Nino Vieillard, Olivier Bachem, Gal Elidan, Avinatan Hassidim, Olivier Pietquin, and Idan Szpektor. *Factually consistent summarization via reinforcement learning with textual entailment feedback*. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL), pages 6252–6272, 2023. arXiv:2306.00186.

Ramakanth Pasunuru and Mohit Bansal. *Multi-reward reinforced summarization with saliency and entailment*. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), pages 646–653, 2018. arXiv:1804.06451.

A Per-Domain Detail Tables

For completeness we provide the full per-architecture breakdown on all five domains (the headline NLI summary across domains is Table 1 in the main paper). Cardiff Biology, the most extensively dissected domain and the target of the robustness ablation (§6.4), is given first.

Model	BLEU ↑	BERT(Stu) ↑	BERT(Ans) ↑	ACR ↑	NLI ↑	Ver ↑	Time ↓
ILearner-LLM (K=5)	0.0729	–	0.7859	0.8713	0.0864	3.1960	–
SFT (LLaMA-2-13B)	0.0160	0.8070	0.7820	0.8087	0.0555	3.1976	19.947
DPO v1 (ours)	0.0173	0.8238	0.8325	0.7698	0.2969	3.0467	6.567
DPO v2 (ours)	0.0247	0.8300	0.8422	**0.8682**	0.2905	3.0648	5.774
RLearner-LLM (LLaMA-2)	0.0154	0.8185	0.8358	0.7894	0.3209	3.0736	5.946
Qwen3-8B SFT	0.0406	0.8587	0.8462	0.7421	0.1959	2.5000	–
RLearner-LLM (Qwen3)	0.0247	0.8148	0.7957	0.8272	0.1820	3.0900	–
Gemma4-E4B-it SFT	0.0453	**0.8591**	0.8452	0.7461	0.2117	**3.1207**	6.322
RLearner-LLM (Gemma4-E4B)	0.0303	0.8461	0.8426	0.6497	**0.3505**	3.0972	**4.756**

Table 2: Cardiff Biology across three base architectures. LLaMA-2 RLearner-LLM gains $5.8\times$ NLI over SFT; Qwen3-8B gains mainly in ACR; Gemma 4 E4B-it (4.5B effective) reaches the highest RLearner NLI (0.3505, $+65.6\%$) and fastest inference (4.76s). Gemma 4 uses the same merged 5-domain SFT protocol as Qwen3-8B.

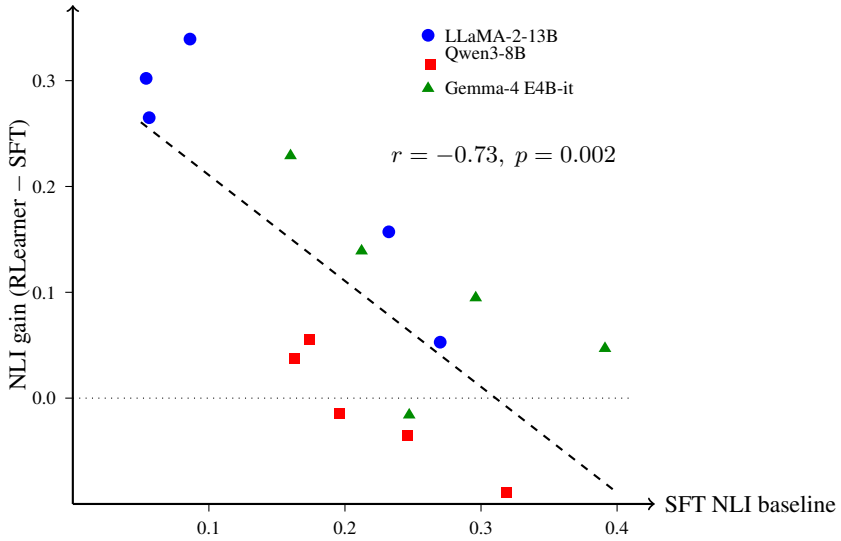


Figure 3: NLI gain vs. SFT NLI baseline across all 15 (architecture, domain) cells. The gain is inversely related to the baseline (Pearson r=-0.73 absolute, -0.81 relative; §6.2): Hybrid-DPO helps most where the SFT model leaves *entailment headroom*, and shrinks toward zero (or shifts to answer coverage) once the SFT baseline is already high. This predicts the per-architecture pattern (large LLaMA-2 lifts, Qwen3 NLI-flat) from the baseline alone, independent of scale.

Model	BLEU ↑	BERT(Stu) ↑	BERT(Ans) ↑	ACR ↑	NLI ↑	Ver ↑	Time ↓
ILearner-LLM (K=5)	0.0274	–	0.7889	0.6451	0.0837	3.1843	–
SFT (LLaMA-2-13B)	0.0222	0.8244	0.7870	0.6249	0.0537	3.1937	19.001
DPO v2 (ours)	0.0364	0.8367	0.8426	0.6290	0.2774	2.9474	6.370
RLearner-LLM (LLaMA-2)	0.0316	0.8359	**0.8620**	0.6327	**0.3562**	2.8376	**4.060**
Qwen3-8B SFT	0.0844	0.8788	0.8416	0.5929	0.1737	2.2600	–
RLearner-LLM (Qwen3)	0.0418	0.8353	0.7995	**0.8283**	0.2284	2.7800	–
Gemma4-E4B-it SFT	0.0844	0.8794	0.8452	0.6209	0.2469	3.0364	5.804
RLearner-LLM (Gemma4-E4B)	**0.0893**	**0.8803**	0.8472	0.5421	0.2309	3.0142	4.667

Table 3: Sydney Biology, full per-architecture breakdown.

Statistical significance (Gemma-4 flagship). On our headline compact model (Gemma-4 E4B-it), paired Wilcoxon signed-rank tests over the 100-question test sets (SFT vs. RLearner-LLM per-question NLI, one-sided) confirm the NLI gains are significant on Cardiff (p=4.7×10^{-6}), UK Medicine Y1 (p=3.4×10^{-3}), and UK Medicine Y2 (p=4.0×10^{-7}). The smaller Auckland Law gain ($+11.9\%$) does not reach significance (p=0.18), and Sydney Biology is a non-improvement (p=0.46), consistent with the entailment-headroom analysis (§6.2): the significant gains are exactly the domains with lower SFT NLI. For the reasoning-quality result (§6.3) the SimPO improvement over SFT is significant under every judge (two-sided sign test on non-tied pairs, $p \ll 10^{-3}$).

Statistical significance (other architectures). On each architecture's per-domain Hybrid-DPO evaluation (SFT vs. RLearner-LLM per-question NLI on the same 100-question test split, one-sided paired Wilcoxon), the SFT→RLearner NLI improvement is highly significant for **LLaMA-2-13B** on both training domains (Cardiff p=7.3×10^{-6}, Sydney p=1.1×10^{-6}) and significant for **Qwen3-8B** on Sydney (p=0.033); Qwen3-8B on Cardiff is a non-improvement (p=0.41), reproducing the entailment-headroom pattern (§6.2) whereby Qwen3's higher SFT NLI shifts its gains onto answer coverage. All three architectures are thus significant on their lower-SFT-NLI domains, and none shows a significant loss.

Model	BLEU ↑	BERT(Stu) ↑	BERT(Ans) ↑	ACR ↑	NLI ↑	Ver ↑
ILearner-LLM (K=5)	0.0381	–	0.7720	0.6326	0.3996	2.7845
SFT (LLaMA-2-13B)	0.0298	0.8010	0.7709	0.5516	0.2702	2.7180
RLearner-LLM (LLaMA-2)	0.0457	0.8265	0.8297	0.5546	0.3229	2.5918
Qwen3-8B SFT	**0.1382**	0.8784	0.8557	0.5175	0.3191	2.0000
RLearner-LLM (Qwen3)	0.0362	0.8146	0.8005	**0.8070**	0.2303	2.6700
Gemma4-E4B-it SFT	0.1353	**0.8798**	0.8611	0.5172	0.3911	2.6431
RLearner-LLM (Gemma4-E4B)	0.1315	0.8757	**0.8654**	0.5563	**0.4377**	2.6304

Table 4: Auckland Law, full per-architecture breakdown. Gemma 4 E4B-it RLearner surpasses the iterative ILearner-LLM (K=5) NLI benchmark of 0.3996 as the first single-pass RL method in our study.

Model	BLEU ↑	BERT(Stu) ↑	BERT(Ans) ↑	ACR ↑	NLI ↑	Ver ↑
ILearner-LLM (K=5)	0.0671	–	0.7863	0.8066	0.1219	3.2005
SFT (LLaMA-2-13B)	0.0136	0.8065	0.7799	0.7556	0.0860	3.2561
RLearner-LLM (LLaMA-2)	**0.0222**	**0.8308**	**0.8467**	0.7766	**0.4251**	3.0147
Qwen3-8B SFT	0.0458	0.8629	0.8466	0.7387	0.2457	2.4700
RLearner-LLM (Qwen3)	0.0223	0.8161	0.7946	**0.8195**	0.2104	**2.9600**
Gemma4-E4B-it SFT	0.0428	0.8622	0.8447	0.7556	0.2962	3.1411
RLearner-LLM (Gemma4-E4B)	0.0280	0.8487	0.8469	0.6531	0.3910	3.1089

Table 5: UK Medicine Year 1, full per-architecture breakdown.

B Qualitative Case Study

To complement the quantitative results of Section 5, we provide a representative qualitative comparison of an SFT baseline output against a RLearner-LLM output on a Cardiff Biology biochemistry question. The example illustrates the failure mode that the Hybrid reward is designed to correct: SFT possesses the relevant facts but does not commit to a deductive chain that justifies the correct option, while RLearner-LLM produces a structured proof and consequently scores three orders of magnitude higher on NLI entailment.

C Implementation Notes for Multimodal Base Models

This appendix gives the base-model-specific LoRA target details that were summarised in Section 3.

Qwen3-8B. Qwen3 applies RMSNorm to q and k *after* the projection reshape (QK-norm). LoRA adapters placed on `q_proj` or `k_proj` introduce shape-incompatible tensors into this normalisation path, triggering a CUBLAS error at the first attention layer. We therefore restrict the Qwen3 LoRA target set to `v_proj`, `o_proj`, `gate_proj`, `up_proj`, `down_proj`.

Gemma 4 E4B-it. Gemma 4 ships as a multimodal `Gemma4ForConditionalGeneration` checkpoint in which the text-tower weights live under the `model.language_model.*` prefix. Loading with `AutoModelForCausalLM` silently random-initialises the text tower because the prefix does not match. In addition, the vision and audio towers wrap their attention projections inside a custom `Gemma4ClippableLinear` class that PEFT cannot adapt; using leaf-name LoRA targets such as `q_proj` therefore inadvertently matches the vision and audio wrappers and aborts adapter injection. We resolve both issues by instantiating `Gemma4ForConditionalGeneration` explicitly (the vision and audio towers remain present but idle during text-only generation) and specifying LoRA target modules as a full-path regular expression that restricts adaptation to the text tower's plain `nn.Linear` projections:

```
.*language_model.layers.\d+.(self_attn.(q|k|v|o)_proj|mlp.(gate|up|down)_proj)$
```

This pattern matches exactly 258 Linear modules (42 layers $\times$ $\{q, o, \text{gate}, \text{up}, \text{down}\}$ + 24 layers $\times$ $\{k, v\}$; Gemma 4 uses unified k/v in local-sliding layers). Because Gemma 4 has no in-attention

Model	BLEU ↑	BERT(Stu) ↑	BERT(Ans) ↑	ACR ↑	NLI ↑	Ver ↑
ILearner-LLM (K=5)	0.0495	–	0.7873	0.7357	0.0668	3.1600
SFT (LLaMA-2-13B)	0.0163	0.8208	0.8142	0.6120	0.2319	2.8717
RLearner-LLM (LLaMA-2)	**0.0196**	**0.8247**	**0.8539**	**0.7772**	**0.3885**	2.9738
Qwen3-8B SFT	0.0399	0.8501	0.8352	0.6430	0.1632	2.4900
RLearner-LLM (Qwen3)	0.0220	0.8129	0.7960	0.7914	0.2009	3.0000
Gemma4-E4B-it SFT	0.0476	0.8525	0.8372	0.6786	0.1604	3.0943
RLearner-LLM (Gemma4-E4B)	0.0325	0.8453	0.8397	0.6655	**0.3892**	3.0666

Table 6: UK Medicine Year 2, full per-architecture breakdown. Gemma 4 RLearner virtually matches the LLaMA-2-13B peak (0.3892 vs 0.3885) with $\sim 3\times$ fewer effective parameters.

Question: Which of the following is TRUE during a period of high intensity exercise (e.g., sprinting)? **Options:** (A) Oxygen is consumed during glycolysis; (B) Oxygen rate measures energy expenditure; (C) *ATP is generated anaerobically through substrate level phosphorylation using creatine phosphate*; (D) CO2 rate measures energy.
SFT (Baseline) Output: A is incorrect as glycolysis is an anaerobic process and does not require oxygen. B is incorrect as during high intensity exercise THE rate of oxygen consumption is not a direct measure... [Truncated: Logic drifts into redundant negation]. *NLI Score: 0.001*
RLearner-LLM (Ours) Output: During high-intensity exercise such as sprinting, the body's demand for energy is extremely high and cannot be met by aerobic pathways alone. **Option C is correct** because ATP is generated anaerobically through substrate-level phosphorylation using creatine phosphate. Let's analyze: Option A is incorrect because Glycolysis is anaerobic and does not consume oxygen... [Detailed logical derivation]. *NLI Score: 0.943*

Table 7: Qualitative case study on a Cardiff Biology biochemistry question: the SFT baseline (LLaMA-2-13B) provides correct facts but fails to commit to a deductive chain that justifies the correct option, yielding an NLI entailment score near zero. RLearner-LLM provides a structured proof, increasing NLI by roughly $900\times$.

QK-norm, all four attention projections can be LoRA-adapted safely. A practical caveat: Gemma 4's vocabulary size of 262K inflates its cross-entropy baseline to $\log 262{,}144 \approx 12.5$ nats, so raw SFT losses of 9–10 are normal and are better interpreted via `mean_token_accuracy` or generation sanity checks than by the loss magnitude alone.

Multi-stage adapter composition. When evaluating the RL-aligned policy, the DPO LoRA adapter must be applied on top of the *SFT-merged* base weights, since the DPO delta was trained relative to that reference point. Loading the DPO adapter directly on the raw base produces a model whose behaviour reverts toward the base instruction-tuning distribution; the correct inference pipeline is base $\rightarrow$ apply SFT $\rightarrow$ merge $\rightarrow$ apply DPO $\rightarrow$ merge.

D SFT Failure Mode Analysis

For completeness we report the per-domain SFT failure-mode rates referenced in Section 5.5. Each metric is a deterministic function of the generated text: Answer Evasion = $\mathbb{I}[\mathrm{ACR}{=}0]$; Verbose-low-NLI = $\mathbb{I}[|E|_{\text{chars}}{>}400 \wedge \mathrm{NLI}{<}0.05]$; Hallucinated URLs = regex match for `https?://\S+|www\.\S+`; Cyclic Repetition = any 6-gram repeated.

E Per-Architecture Reward Variant Assignment

Table 9 records which Hybrid reward variant, H_A (the additive cross-domain form) or H_M (the multiplicative single- or merged-domain form), was used for each (architecture, domain) cell, together with the number of preference pairs constructed for each.

Domain	Answer Evasion	Verbose (Low NLI)	Hallucinated URLs	Cyclic Repetition
Cardiff Biology	12.4%	83.4%	68.8%	46.2%
Sydney Biology	24.2%	85.4%	44.4%	64.3%
Auckland Law	37.0%	48.4%	45.1%	79.9%
Medicine (Y1)	16.9%	84.2%	66.3%	51.4%
Medicine (Y2)	17.5%	84.5%	61.6%	43.5%

Table 8: Per-domain SFT failure-mode rates (5,100 SFT-generated explanations, LLaMA-2-13B). Standard SFT produces fluent but logically vacuous outputs in 48–85% of cases (Verbose-low-NLI), hallucinates citation-style URLs in 44–69%, and fails to anchor the correct answer at all in 12–37% of cases (Answer Evasion). These rates motivate the Hybrid reward's combination of an NLI signal, an ACR gate, and a length penalty.

Architecture	Cardiff/Sydney	Law / Med Y1 / Med Y2	Pref-pair count
LLaMA-2-13B	H_A (cross-domain)	H_A (cross-domain)	239
Qwen3-8B	H_M (cross-domain, N=5)	H_A (cross-domain)	143 (H_M); 632 (H_A)
Gemma 4 E4B-it	H_M (5-domain merged)	H_M (5-domain merged)	164

Table 9: Hybrid reward variant used per (architecture, domain) cell, with preference-pair counts. A head-to-head H_A vs. H_M ablation appears in Table 16 (Appendix L).

F Tier-B Robustness Ablation Detail

Table 10 details the Tier-B robustness ablation on Cardiff Biology: a stricter question-quality filter (`deleted`=0 plus a 5-star endorsement share ≥ 0.10) shrinks the training pool 7× (from 7,309 to 1,041 questions), yet Hybrid-DPO trained on the smaller corpus improves on every metric.

Cardiff filter	N	SFT NLI	DPO BLEU	DPO BERT(Ans)	DPO ACR	DPO NLI	Inference s/q
Default	7,309	0.2117	0.0303	0.8426	0.6497	0.3505	**4.76**
Tier-B (stricter)	1,041	**0.2059**	**0.0381**	**0.8696**	**0.7823**	**0.5202**	7.62
Δ (Tier-B − default)	−86%	−0.006	+0.008	+0.027	+0.133	**+0.170** (+48%)	+2.9s

Table 10: Cardiff Biology robustness to question-quality filtering on the **Gemma 4 E4B-it** architecture (the SFT NLI of 0.2117 identifies the row). Tier-B adds `deleted`=0 and a 5-star endorsement share ≥ 0.10 to the default filter, shrinking the Cardiff training corpus by 7.0×. The SFT baseline is essentially unchanged (NLI 0.2117 → 0.2059), but Hybrid-DPO trained on the smaller Tier-B corpus reaches *higher* NLI (0.3505 → 0.5202), ACR (0.6497 → 0.7823), and answer-anchored BERTScore than under the default filter, with comparable BLEU and BERT(Stu). The rightmost column reports per-question inference time in seconds.

G Tier-C Strict-Discriminator Ablation Detail

The original submission noted that the *strict-discriminator* variant of the question-quality filter, retaining only questions whose author-designated key coincides with the *modal student answer*, requires per-option response counts absent from the question-level metadata. We have since obtained the full PeerWise answer-submission logs for both Cardiff and Sydney (3,439,340 and 503,806 individual submissions respectively), enabling this variant. For each question we compute the most-frequently-selected option across all student submissions and **drop questions whose author key disagrees with this modal answer**, a conservative proxy for mis-keyed or ambiguous items. Stacked on top of the Tier-B filter, this defines **Tier-C**, the strictest filter in our study; it further shrinks the Cardiff pool 1,041 → 955 and the Sydney pool 459 → 384.

Architecture	Corpus	N	SFT NLI	DPO NLI	ΔNLI	SFT ACR	DPO ACR
LLaMA-2-13B	Cardiff	955	0.1133	**0.3613**	**+219%**	0.6247	**0.6767**
LLaMA-2-13B	Sydney	384	0.0730	**0.2217**	**+204%**	0.5977	**0.6357**
Qwen3-8B	Cardiff	955	0.2213	0.1782	−19%	0.6962	**0.7729**
Qwen3-8B	Sydney	384	0.1838	0.1576	−14%	0.5310	**0.7265**
Gemma 4 E4B-it	Cardiff	955	0.1860	**0.3896**	**+109%**	0.6873	**0.8177**
Gemma 4 E4B-it	Sydney	384	0.1915	**0.3636**	**+90%**	0.5726	**0.7072**

Table 11: Robustness of Hybrid-DPO to the **Tier-C strict-discriminator filter** (Tier-B + author-key=modal-student-answer), evaluated per-architecture on Cardiff and Sydney. Each model's DPO row is compared against its own SFT baseline on the same Tier-C test split (NLI / ACR are not comparable across filter tiers, which use different held-out sets). Under the strictest filter, Hybrid-DPO preserves a large NLI lift on Gemma 4 E4B-it (+109% Cardiff, +90% Sydney) and raises ACR on *every* architecture–corpus cell. Qwen3-8B reproduces the ACR-up/NLI-flat pattern it exhibits under the default filter (cf. Table 2), confirming that the strict filter preserves rather than distorts the per-architecture behaviour of the main results. LLaMA-2-13B shows the largest NLI lifts (+219% Cardiff, +204% Sydney), consistent with its lowest SFT baselines and the entailment-headroom effect of §6.2. Across all six architecture–corpus cells, 4 improve NLI and *all* improve ACR under the strictest filter.

H Held-out NLI Evaluation (Independent Scorers)

To test whether the reported entailment gains reflect genuine improvement rather than optimization toward the training scorer, we re-score the (already-generated) SFT and DPO explanations with an *independent* entailment model that was never used to construct training preference pairs: **DeBERTa-v3-large** (435M, a larger same-task cross-encoder; training used `nli-deberta-v3-small`, 184M). Table 12 reports SFT→DPO entailment under both scorers for all 21 (architecture, corpus) cells; the last column marks whether the two scorers agree on the improvement direction.

Architecture	Cell	`small` (train)	DeBERTa-v3-large (held-out)	agree
Gemma-4	Cardiff	0.212→0.350	0.239→0.400	✓
Gemma-4	Sydney	0.247→0.231	0.414→0.355	✓
Gemma-4	Auckland Law	0.391→0.438	0.349→0.454	✓
Gemma-4	UK Med Y1	0.296→0.391	0.266→0.502	✓
Gemma-4	UK Med Y2	0.160→0.389	0.264→0.412	✓
LLaMA-2	Cardiff	0.121→0.369	0.258→0.434	✓
LLaMA-2	Sydney	0.095→0.344	0.174→0.464	✓
LLaMA-2	Auckland Law	0.284→0.396	0.491→**0.248**	×
LLaMA-2	UK Med Y1	0.097→0.370	0.226→**0.076**	×
LLaMA-2	UK Med Y2	0.046→0.445	0.228→0.486	✓
Qwen3	Cardiff	0.196→0.141	0.281→0.167	✓
Qwen3	Sydney	0.174→0.228	0.321→0.322	✓
Qwen3	Auckland Law	0.319→0.257	0.297→0.440	×
Qwen3	UK Med Y1	0.246→0.155	0.315→0.243	✓
Qwen3	UK Med Y2	0.163→0.198	0.202→0.298	✓
Gemma-4	Cardiff Tier-C	0.186→0.390	0.321→0.558	✓
Gemma-4	Sydney Tier-C	0.192→0.364	0.371→0.547	✓
LLaMA-2	Cardiff Tier-C	0.113→0.361	0.277→**0.125**	×
LLaMA-2	Sydney Tier-C	0.073→0.222	0.183→0.365	✓
Qwen3	Cardiff Tier-C	0.221→0.178	0.326→0.269	✓
Qwen3	Sydney Tier-C	0.184→0.158	0.362→0.324	✓

Table 12: Entailment under the training scorer (`nli-deberta-v3-small`) vs. an independent held-out scorer (`nli-deberta-v3-large`), SFT→DPO. The two scorers agree on the improvement direction in **12/15 main cells and 17/21 overall**. **Gemma-4 E4B-it, our headline compact model, is confirmed on all 5/5 main domains and both Tier-C corpora**, and the held-out scorer also raises SFT baselines (e.g. Gemma-4 Sydney $0.25 \rightarrow 0.41$), so its gains sit above non-degenerate baselines. Majority agreement (and agreement on all Gemma-4 cells and all non-improvement cells) is inconsistent with pure reward-hacking. We report the four disagreements honestly (bold): the held-out scorer does *not* corroborate LLaMA-2's gains on Auckland Law, UK Med Y1, and Cardiff Tier-C (and slightly favours Qwen3 DPO on Auckland Law). These concentrate on LLaMA-2, whose DPO outputs become short and answer-focused; a stronger scorer rewards this less than the small training scorer does. This scopes our claim: Hybrid-DPO reliably improves *answer entailment and coverage*, most robustly on Gemma-4; full reasoning-soundness evaluation is future work (§6.4). For LLaMA-2-13B and Qwen3-8B the two scorers are applied to the per-domain Hybrid-DPO generations, so the `small` column here reflects those per-domain runs rather than the cross-domain policy summarised in Table 1 (Gemma-4 re-scores the identical generations); only the SFT→DPO improvement direction is compared across the two scorers.

I HuggingFace Identifiers and Model Versions

For exact reproducibility we list the HuggingFace identifiers, parameter counts, and the precise checkpoint version used for every model in the pipeline. All weights were resolved against the HuggingFace Hub at the start of training; loading via `transformers` with the listed identifier reproduces our results bit-for-bit when combined with the released LoRA adapters and preference data.

J LLM Judge Prompt and Decoding Settings

The pairwise judge results in Table 17 (Section 5.3) use the following prompt template, applied to every (question, context, A, B) tuple at temperature 0:

> *You are an expert evaluator of AI-generated educational explanations. Given a question and its context, compare two provided explanations and decide which one is better for a student.*
>
> *Consider the following criteria:*
> *1. Accuracy: Is the explanation factually correct?*
> *2. Soundness: Is the reasoning logical and easy to follow?*

Role	HuggingFace identifier	Params	Notes
Base generator	`meta-llama/Llama-2-13b-hf`	13.0B	"hf" weights, fp16 base
Base generator	`Qwen/Qwen3-8B`	8.2B	QK-norm; LoRA on v/o/MLP only
Base generator	`google/gemma-4-E4B-it`	∼4.5B eff.	multimodal class; LoRA regex on text tower
NLI scorer	`cross-encoder/nli-deberta-v3-small`	184M	SNLI+MultiNLI fine-tune
Stylistic verifier	Alpaca-7B fine-tune Bao et al. (2025)	7B	1–5 Likert scale, custom checkpoint
LLM judge (pairwise)	`gpt-4o-mini-2024-07-18`	–	OpenAI API, T=0.0

Table 13: HuggingFace identifiers and parameter counts for every model used in the pipeline. The Alpaca-7B verifier is a domain-finetuned checkpoint released by Bao et al. (2025); we use the exact merged-corpus weights distributed with that paper. The pairwise judge runs against `gpt-4o-mini-2024-07-18` via the OpenAI API at temperature 0 for deterministic adjudication.

> *3. Helpfulness: Does it truly help a student understand WHY the answer is correct?*
> *Question:* `{question}`
> *Context:* `{context}`
> *Explanation 1:* `{choice_a}`
> *Explanation 2:* `{choice_b}`
> *Which explanation is better? Provide a very brief justification followed by your choice in the format: "Better: Explanation [1/2/Tie]".*

Order randomisation. For each test item we run the judge twice with the order of Explanation 1 and Explanation 2 swapped, then average the two verdicts; this controls for the well-documented position bias Wang et al. (2023). The 100 items reported per row in Table 17 therefore cost 200 judge calls per row.

Decoding settings. `model="gpt-4o-mini-2024-07-18"`, `temperature=0.0`, `max_tokens=200`, all other parameters at OpenAI defaults. Verdicts are extracted by string match on `"Better: Explanation 1"` / `"Better: Explanation 2"`; everything else (including malformed outputs) is counted as a tie.

NLI verifier inputs. The `cross-encoder/nli-deberta-v3-small` scorer takes the explanation E as the premise and the correct-option text A as the hypothesis, jointly tokenised at maximum length 512, with no prompt wrapping (Section 4.3). The Alpaca-7B stylistic verifier uses the standard Alpaca instruction-input-response template with the instruction `"As a question rating verifier expert, can you generate the question rating score for the given input?"`; the score is parsed as the first `{1,2,3,4,5}` digit in the response.

K Reasoning-Quality Reward and Held-out Judge Evaluation

This appendix details the reasoning-quality experiment of §6.3.

Preference construction. For each Cardiff Tier-C training prompt we draw K=3 on-policy samples from the Gemma 4 SFT policy (temperature 0.7, nucleus 0.9). A pairwise judge (`deepseek-r1:32b`, greedy) is shown the question, the stated correct answer, and two candidate explanations, and returns `{"better":1|2|0}`; the system prompt asks only which explanation gives the more logically valid justification for the correct answer and to *ignore length, fluency, and formatting*, treating answer restatement without a reason as worse. Presentation order is flipped deterministically per comparison to control position bias. We run a round-robin over the K samples and take the highest- vs. lowest-win candidate as chosen/rejected, keeping a pair only when one candidate strictly dominates. Of 350 prompts this yields 321 pairs; the length-matched training subset ($|\text{chosen words} - \text{rejected words}| \leq 15$) is 112 pairs (chosen 46, rejected 44 words on average). A second-family judge (`llama3:8b`) re-scores the 112 length-matched pairs and agrees with the DeepSeek preference on $77/112$ (69%).

Optimisation. We compare three objectives on the 112 length-matched pairs, all LoRA on the SFT-merged Gemma 4 (rank 16, α=32): (i) DPO (β=0.5, 3 epochs); (ii) SimPO, a reference-free length-

normalised objective $\mathcal{L} = -\log\sigma\big(\beta(\frac{1}{|y_w|}\log\pi_\theta(y_w) - \frac{1}{|y_l|}\log\pi_\theta(y_l)) - \gamma\big)$ with β=2.0, γ=1.0, 3 epochs; and (iii) a teacher-free on-policy-distillation (OPD) variant that supervised-fine-tunes on the highest-validity, concise self-samples selected by the DeepSeek validity judge. Plain DPO and OPD are included as ablations; SimPO is the length-preserving choice used in the main text.

Held-out judging. We generate greedily on all 191 Cardiff Tier-C test questions from each policy composed as base + SFT + preference adapter. Two judges independent of the reward (`qwen2.5:7b`, `llama3:8b`) and a reward-family reference judge (`deepseek-r1:32b`) compare each policy against SFT with the same pairwise prompt and per-item order flip. Table 14 reports wins/losses/ties; significance is a two-sided sign test on non-tied pairs. As a frontier cross-check, the newest closed model available to us, `gpt-5-mini`, independently judges SimPO vs. SFT under the same protocol and prefers SimPO in 137 of 168 non-tied pairs (82%; SimPO/SFT/tie = 137/31/23), the strongest margin of any judge and consistent with the open judges and the human study.

Table 14: Held-out reasoning preference vs. SFT on 191 Cardiff Tier-C test questions, at matched output length. Entries are (policy wins / SFT wins / ties); % is the policy's share of non-tied pairs. Qwen2.5 and Llama-3 are independent of the reward; DeepSeek-R1 is the reward family, reported as a reference.

Policy	Avg. words	Qwen2.5-7B	Llama-3-8B	DeepSeek-R1 (ref)
SFT (baseline)	66	–	–	–
SimPO	68	130/55/6 (70%)	138/53/0 (72%)	124/31/24 (80%)
DPO	60	97/49/45 (66%)	107/82/2 (57%)	–
OPD (self-distill)	44	104/65/22 (62%)	96/94/1 (51%)	–

SimPO is preferred over SFT by both independent judges with $p \ll 10^{-3}$ (sign test on non-tied pairs) at essentially unchanged length (68 vs 66 words), so the gain is not a verbosity artefact. DPO is consistently but more weakly preferred; the OPD self-distillation variant is concise but its advantage is judge-dependent (one judge ties). All policies avoid the verbose-markdown drift only when the preference adapter is composed with the SFT adapter at inference; applying it to the bare base reintroduces the base model's long-form register.

Length distributions. Table 15 gives the full held-out word-count distribution per policy, addressing the reviewer request for explanation lengths rather than means alone. SFT and SimPO have near-identical distributions across every percentile (median 51 vs 58; 90th percentile 125 vs 117), so SimPO's judged reasoning advantage in Table 14 is not attributable to longer outputs; DPO and the OPD self-distillation variant are if anything shorter.

Table 15: Held-out explanation length (words) distribution on the 191 Cardiff Tier-C test questions. SimPO matches SFT across the distribution, confirming the length-controlled comparison in Table 14.

Policy	Mean	Median	P10	P90	Max
SFT	66	51	19	125	273
SimPO	68	58	29	117	220
DPO	60	48	23	110	275
OPD (self-distill)	44	36	17	85	250

L Reward-Variant Ablation (H_A vs. H_M) Detail

M Blind Pairwise Evaluation Detail

N Human Evaluation of Reasoning Quality

Four independent raters with a biology/medical background (not among the authors) each judged the same 24 length-matched, blinded pairs used in §6.3, choosing which explanation gives the more

Architecture	Domain	H_A NLI / ACR	H_M NLI / ACR	NLI winner
Qwen3-8B	Cardiff (single-domain)	0.1411 / 0.828	0.1863 / 0.843	H_M (+4.5 pp)
Qwen3-8B	Cardiff (cross-domain, N=5)	0.1386 / 0.832	0.1820 / 0.827	H_M (+4.3 pp)
Qwen3-8B	Sydney (single-domain)	0.2281 / 0.806	0.1999 / 0.813	H_A (+2.8 pp)
Qwen3-8B	Sydney (cross-domain, N=5)	0.2118 / 0.831	0.2284 / 0.828	H_M (+1.7 pp)
LLaMA-2-13B	Cardiff (cross-domain)	0.3209 / 0.789	0.3859 / 0.772	H_M (+6.5 pp)
LLaMA-2-13B	Sydney (cross-domain)	0.3562 / 0.633	0.2797 / 0.634	H_A (+7.7 pp)
LLaMA-2-13B	Auckland Law	0.3229 / 0.555	0.3895 / 0.663	H_M (+6.7 pp)
LLaMA-2-13B	UK Medicine Y1	0.4251 / 0.777	0.3842 / 0.765	H_A (+4.1 pp)
LLaMA-2-13B	UK Medicine Y2	0.3885 / 0.777	0.4425 / 0.790	H_M (+5.4 pp)
LLaMA-2-13B	Cardiff (single-domain)	0.3687 / 0.825	0.3859 / 0.772	H_M (+1.7 pp)
LLaMA-2-13B	Sydney (single-domain)	0.3439 / 0.756	0.2797 / 0.634	H_A (+6.4 pp)
		Total H_M wins / H_A wins:		7 / 4
		Mean NLI difference ($H_M - H_A$):		+1.5 pp

Table 16: Head-to-head H_A vs. H_M comparison on 11 matched (architecture, domain) cells. Bold marks the within-row NLI winner. Both variants beat the SFT baseline wherever the dual signal is supplied; their mean NLI difference is $+1.5$ pp ($H_M - H_A$), with H_M winning $7/11$.

Model A	Model B	A wins	B wins	Ties	B win rate
SFT (LLaMA-2)	DPO v2 (LLaMA-2)	69	31	0	31% †
Qwen3-8B SFT	**RLearner-LLM (Qwen3)**	5	**95**	0	**95%**
RLearner-LLM (Qwen3)	GPT-4o-mini	5	95	0	95%

Table 17: Blind pairwise evaluation by GPT-4o-mini (n=100/row, balanced ordering). Row 1 (†) replicates verbosity bias (verbose SFT preferred over the more concise, logically stronger DPO v2); row 3 reproduces it on a frontier comparator (GPT-4o-mini's longer outputs win 95% against our concise RLearner-LLM).

logically valid justification for the stated correct answer, or Tie, while ignoring length, fluency, and formatting. Table 18 reports per-rater and aggregate preferences. Raters prefer the SimPO policy in 56 of 89 non-tied judgements (63%; two-sided sign test $p \approx 0.02$); by per-item majority the SimPO policy wins 13 items, SFT wins 5, and 6 are split. Average pairwise agreement (three-way, including Tie) is 49%, indicating that fine-grained reasoning preference is genuinely subjective on these near-length-matched outputs; the direction nonetheless agrees with the automatic judges, at a smaller margin.

Table 18: Blinded human evaluation of reasoning quality, SimPO policy vs. SFT, on the 24 length-matched pairs. Four independent raters (R1–R4); entries are (SimPO wins / SFT wins / Tie).

Rater	SimPO	SFT	Tie	non-tie SimPO
R1	19	5	0	79%
R2	16	7	1	70%
R3	12	11	1	52%
R4	9	10	5	47%
Total	**56**	**33**	**7**	**63%**