On-Policy Distillation Meets Off-Policy GRPO:
Training Sparse Instruction-Following Rerankers that Beat Their Teachers

Anonymous ACL submission

Abstract

Compact instruction-following rerankers are attractive for deployment, but existing distillation pipelines typically train students by offline imitation of teacher outputs on a fixed set of examples, constraining the student to the teacher's observed ranking space. We revisit reranker distillation through the lens of reinforcement learning.

We propose a two-stage framework combining off-policy teacher optimization with on-policy student distillation. In Stage 1, a 4B teacher reranker is trained with off-policy GRPO using LLM-judge feedback on 88K instruction-following examples. In Stage 2, a compact 1B student is distilled via on-policy GRPO: the student samples rankings from its own policy and receives soft teacher-derived rewards, coupling exploration with knowledge transfer.

Our strongest evidence appears on out-of-distribution evaluation. On the MAIR benchmark, the proposed student reaches 0.767 nDCG@6, outperforming offline knowledge distillation by +4.6 points under the same supervision budget. Controlled ablations show that replacing student-driven on-policy exploration with teacher-driven sampling reduces nDCG@6 by 3.1 points, and that distilling on-policy from an unaligned teacher recovers only part of the gain, indicating that teacher strengthening and on-policy student learning are complementary. Replacing soft rewards with hard labels improves in-distribution fit but degrades OOD performance by 3.1 points, revealing a clear accuracy–generalization tradeoff. On the combined validation benchmark covering 9,861 queries across eight datasets, the resulting 1B reranker achieves 0.762 nDCG@6, outperforming Cohere Rerank v3.5 and Cohere Rerank v4.0-fast while running up to 4× faster than 4B alternatives.

1 Introduction

Neural rerankers are now a core component of modern retrieval systems, providing fine-grained relevance estimation beyond first-stage retrieval pipelines (Nogueira and Cho, 2019; Khattab and Zaharia, 2020; Yates et al., 2021). Recent work extends reranking to *instruction-following* settings, where users specify ranking criteria in natural language rather than through implicit query intent alone (Asai et al., 2023; Weller et al., 2024a; Oh et al., 2024). This capability is especially important in enterprise and assistant-facing retrieval, where ranking decisions may depend on temporal constraints, source preferences, audience requirements, or other multi-factor instructions.

Despite this progress, training *compact* instruction-following rerankers that generalize robustly remains difficult. Large proprietary rerankers offer strong performance but are costly and latency-sensitive in production. More fundamentally, existing student training approaches typically reduce distillation to offline imitation: the student is trained to match teacher-provided labels or score distributions on a fixed set of examples, constraining the student to the teacher's observed ranking space and potentially transferring teacher-specific biases under distribution shift.

We revisit reranker distillation from a reinforcement learning perspective. Our central hypothesis is that effective capability transfer requires *student-driven exploration*: the student should generate rankings from its own policy and learn from teacher-derived rewards on those rankings. We propose a two-stage pipeline: **Stage 1** trains a 4B teacher using off-policy GRPO with LLM-judge feedback; **Stage 2** distills into a compact 1B reranker using on-policy GRPO with soft teacher rewards.

On-policy distillation has recently emerged as a strong post-training paradigm for reasoning and code generation, where the student receives teacher supervision on its own generated token trajectories (Agarwal et al., 2024; Lu and Thinking Machines Lab, 2025; Qwen Team, 2025). We carry

this paradigm into a different action space: *rankings* rather than token sequences. A ranking is a permutation over candidate documents, and teacher supervision is a soft reward over sampled rankings rather than a logit distribution over next tokens.

Our contributions:

- **Mechanism finding.** We identify *where* teacher supervision is evaluated as the dominant factor: on-policy and off-policy variants share teacher-derived supervision but differ in whose policy generates rankings, and the on-policy variant improves nDCG@6 by 3.1 points on MAIR with non-overlapping CIs.
- **Teacher strengthening as precondition.** Distilling on-policy from an unaligned teacher (A8) improves over offline KD but remains 2.6 points below the full pipeline on MAIR, showing both stages are complementary.
- **Soft rewards as regularization.** Hard labels improve in-distribution fit but degrade OOD performance by 3.1 points, supporting a regularization interpretation of soft teacher rewards.
- **Deployable artifact.** The resulting 1B reranker matches or exceeds Cohere Rerank v3.5/v4.0-fast, BGE v2 M3, and Jina v2 while running roughly 3× faster than 4B alternatives.

2 Related Work

Instruction Following Reranking. Neural rerankers improve retrieval quality by rescoring candidate documents, often using cross-encoder architectures (Nogueira and Cho, 2019; Yates et al., 2021). Recent work extends this to instruction-following settings where models condition on natural-language ranking criteria (Asai et al., 2023; Weller et al., 2024a; Oh et al., 2024), supported by benchmarks such as FOLLOWIR and INSTRUCTIR. Systems such as ZeRank-2 (ZeroEntropy, 2025) demonstrate that instruction-following reranking benefits from RL-based alignment, while compact open rerankers such as Llama-Nemotron-Rerank-1B-v2 (NVIDIA, 2025) provide strong efficiency-oriented starting points. We build on this setting but ask how a compact reranker should acquire instruction-following behavior from a stronger teacher without inheriting the teacher's fixed prediction distribution.

Prompt-based and LLM-based Reranking. Promptriever (Weller et al., 2024b) makes dense retrieval models instruction-sensitive through prompting. RankGPT and related methods treat ranking as an inference-time prompting problem (Qin et al., 2024; Ma et al., 2023), while RankVicuna and RankZephyr show that open-source LLMs can be effective zero-shot listwise rerankers (Pradeep et al., 2023a,b). We differ in goal: rather than large prompted LLMs at inference time, we train a compact 1B reranker through a two-stage RL pipeline.

Reinforcement Learning for Ranking. RL has become standard for optimizing non-differentiable objectives in language generation and alignment (Ouyang et al., 2022; Rafailov et al., 2023). GRPO (Shao et al., 2024) provides efficient policy-gradient optimization via group-normalized advantages. Prior work in offline RL emphasizes that policies trained on fixed data can be limited by distributional coverage, whereas on-policy interaction can correct errors on states induced by the current policy (Levine et al., 2020). In reranking, RL has mainly been used as an alignment stage for large rerankers or in off-policy settings. Our method uses GRPO in two distinct roles: off-policy for the teacher and on-policy for student distillation.

Knowledge Distillation for Retrieval. Knowledge distillation transfers behavior from large teachers to smaller students (Hinton et al., 2015). In retrieval, prior work commonly performs offline distillation from fixed teacher scores (Hofstätter et al., 2021). Born-again distillation has shown that students can sometimes outperform teachers via a regularization effect from soft supervision (Furlanello et al., 2018). LLM-as-judge methods (Zheng et al., 2023; Dubois et al., 2024) provide scalable soft reward signals richer than binary labels. We depart from standard offline KD by treating distillation as policy learning: the student samples rankings from its own policy and receives teacher-derived rewards, enabling exploration that offline KD does not.

3 Method

Model naming conventions. **ZeRank-2** denotes the 4B teacher backbone (ZeroEntropy, 2025). **Teacher GRPO** refers to ZeRank-2 after Stage 1 off-policy RL. **Base-1B** denotes the Llama-Nemotron-Rerank-1B-v2 student backbone (NVIDIA, 2025), and **Distilled-1B** denotes

the final student after Stage 2 on-policy distillation.

3.1 Problem Setup

Given a query q, instruction i, and candidate document set $D = \{d_1, \ldots, d_n\}$, the reranker defines a scoring function over documents and induces a distribution over rankings. Our training pipeline has two stages (Figure 1). Stage 1 trains a stronger teacher using off-policy GRPO with LLM-judge feedback. Stage 2 distills the teacher into a compact student using on-policy GRPO, where the student samples rankings from its own policy and receives teacher-derived soft rewards. The central difference from offline distillation is that supervision is generated on the student's sampled rankings rather than on a fixed teacher-generated distribution.

3.2 Stage 1: Off-Policy GRPO Teacher Training

For each training example (q, i, D), the teacher induces a distribution over rankings and we sample candidate rankings $\pi^{(1)}, \ldots, \pi^{(k)}$ using Plackett–Luce sampling. An LLM judge assigns a reward $r(\pi^{(j)})$ to each sampled ranking. Following GRPO (Shao et al., 2024), we compute group-normalized advantages

$$A(\pi^{(j)}) = \frac{r(\pi^{(j)}) - \mu}{\sigma}, \tag{1}$$

where μ, σ are the mean and standard deviation of the sampled rewards. The teacher is updated with

$$\begin{aligned}\mathcal{L}_{\text{teacher}} = & -\mathbb{E}_{\pi \sim p_\theta}[A(\pi) \log p_\theta(\pi)] \\ & + \beta_{\text{KL}} \text{KL}(p_\theta \| p_{\text{ref}}) - \alpha_{\text{ent}} H(p_\theta).\end{aligned} \tag{2}$$

3.3 Stage 2: On-Policy GRPO Distillation

In Stage 2, the student samples rankings from its *own* policy: $\pi^{(1)}, \ldots, \pi^{(k)} \sim p_{\theta_S}$. The teacher assigns a soft reward $r_T(\pi)$ to each sampled ranking, yielding student-specific supervision. We compute

$$A_T(\pi^{(j)}) = \frac{r_T(\pi^{(j)}) - \mu_T}{\sigma_T}, \tag{3}$$

and optimize the student with

$$\begin{aligned}\mathcal{L}_{\text{student}} = & -\mathbb{E}_{\pi \sim p_{\theta_S}}[A_T(\pi) \log p_{\theta_S}(\pi)] \\ & + \lambda_{\text{KL}} \text{KL}(p_{\theta_S} \| p_T) - \alpha_{\text{ent}} H(p_{\theta_S}).\end{aligned} \tag{4}$$

The key property is that supervision is evaluated on the student's sampled rankings rather than on teacher-specified rankings, making distillation an on-policy learning process.

Dataset	Train	Val
FollowIR (TREC)	445	49
InstructIR	8,915	991
InfoSearch	4,296	477
InfIR (MS MARCO)	34,883	3,876
InfIR (MetaMath)	6,394	710
InfIR (LeetCode)	2,286	254
InfIR (Robust04)	1,765	196
M-BEIR (WebQA)	29,771	3,308
Total	**88,755**	**9,861**

Table 1: Instruction-following reranking data used for training and validation.

3.4 On-Policy Distillation and Generalization

Offline distillation supervises the student only on a fixed teacher distribution, encouraging imitation of rankings the teacher already prefers. On-policy distillation instead evaluates teacher-derived supervision on rankings sampled from the student's own policy, expanding the effective coverage of the training signal. We therefore expect the gap between offline KD and on-policy distillation to be smallest in-distribution and larger under distribution shift—exactly the pattern observed in our experiments.

3.5 Objective Components

Our ablations show that the dominant learning signal comes from policy-gradient optimization with teacher rewards. The KL and entropy terms are optional stabilizers with limited effect on in-distribution performance but can improve OOD robustness. We retain them in the final model and analyze their effect in Section 6.

4 Experimental Setup

4.1 Datasets

We evaluate on a diverse instruction-following reranking benchmark constructed from eight datasets spanning web search, code, math, news, and multi-hop retrieval. The training split contains 88,755 examples and the validation set contains 9,861 queries (Table 1).

We evaluate on MAIR subsets topically relevant to enterprise-style queries over heterogeneous corpora, spanning ad hoc retrieval (Core_2017, DD_2016), FAQ matching (Quora, CQADupStack), instruction-sensitive selection (IFEval), scientific/biomedical evidence retrieval (SciFact, SciDocs, Trec-Covid, NFCorpus, LitSearch), financial retrieval (FiQA), and argumentative retrieval (ArguAna, Touche). These subsets are representative

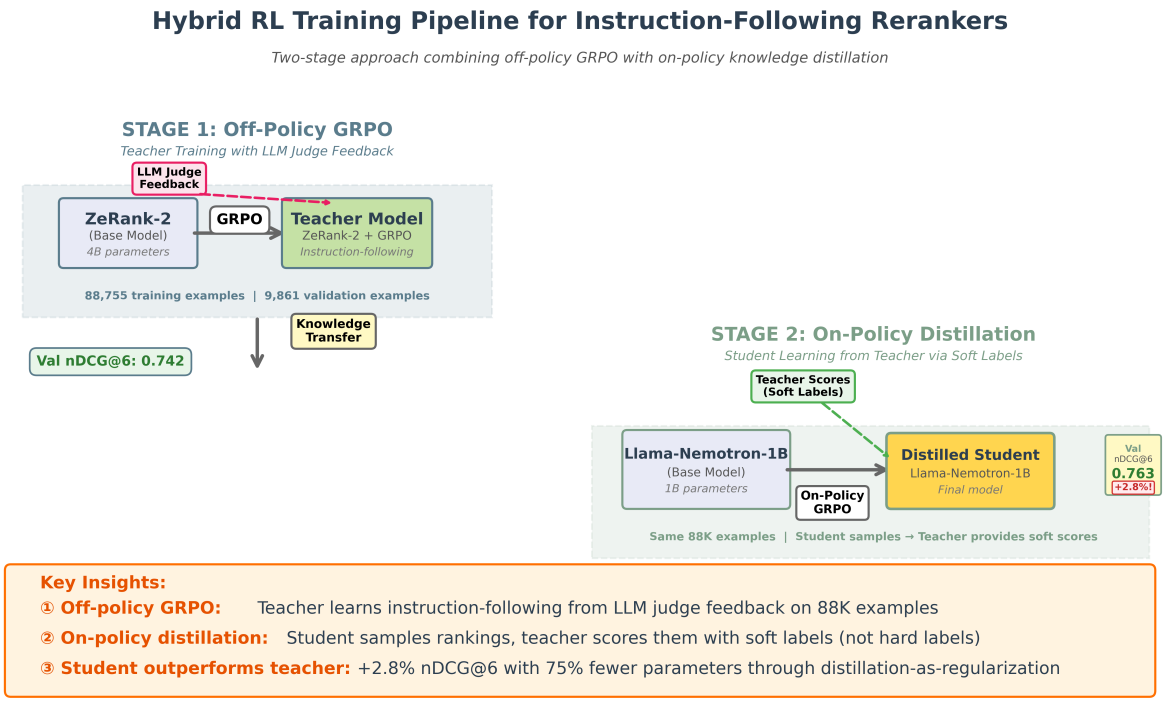


Figure 1: Two-stage training pipeline. Stage 1 trains a 4B teacher with off-policy GRPO and LLM-judge feedback. Stage 2 distills into a 1B student via on-policy GRPO, where the student samples rankings from its own policy and receives soft teacher rewards.

of our target deployment setting rather than being chosen to optimize results on any single domain.

4.2 Models and Baselines

Our teacher model is **ZeRank-2** (ZeroEntropy, 2025) (4B). Our student backbone is **Llama-Nemotron-Rerank-1B-v2** (NVIDIA, 2025). We compare against: (i) internal references (Base-1B, ZeRank-2 base, Teacher GRPO); (ii) training ablations including supervised BCE, offline KD, off-policy GRPO, on-policy GRPO, distillation from unaligned base teacher (A8), objective-component removals, and hyperparameter sweeps; (iii) external baselines (BGE Reranker v2 M3, Jina Reranker v2, Qwen3-Reranker-4B, Cohere Rerank v3.5, Cohere Rerank v4.0-fast). We exclude Jina, BGE, and Cohere from MAIR due to potential train–test contamination.

4.3 Training Configuration

Stage 1. Off-policy GRPO with LLM-judge rewards: group size $k = 6$, learning rate 5×10^{-6} with 3% warmup, batch size 64, 10 epochs.

Stage 2. On-policy GRPO with teacher-derived soft rewards: group size $k = 6$, temperature 1.0, KL weight 1.0, entropy coefficient 0.01, learning rate 2×10^{-6} with 3% warmup, batch size 64, 3 epochs. Set B varies these hyperparameters. All training on NVIDIA H200 GPUs.

LLM Judge Validation. Because Stage 1 relies on judge-derived reward signals, it is important to verify that the supervision is directionally consistent across strong LLM evaluators. We therefore compare five judges (Gemini 2.5, Claude Sonnet 4, GPT-4o, Mistral, and OSS-20B) on criteria directly relevant to instruction-following reranking.

Figure 2 shows that inter-judge consistency is strong on two core dimensions. For relevance, judges exhibit high Spearman rank correlation, indicating that they largely preserve the same ordering of candidate passages. For answer containment, pairwise agreement is also high, suggesting that judges are broadly consistent in identifying whether a passage contains the required answer signal. These results support our use of judge-derived outputs as *soft* reward supervision for teacher training, while still treating them as an approximation rather than a substitute for human annotation.

Beyond inter-judge consistency, we also validate the Stage 1 judge against human annotations on a manually labeled relevance set. The LLM judge achieves 92.6% agreement with human labels (174/188 examples), with Cohen's $\kappa = 0.84$, indicating near-perfect agreement. This result suggests

that the judge is not merely self-consistent across strong LLM evaluators, but also closely aligned with human relevance judgments. The disagreement pattern is asymmetric: most errors are false negatives, indicating that the judge is somewhat stricter than human annotators rather than overly permissive.

4.4 Evaluation Metrics

We report **nDCG@6** and **MRR@6** as primary ranking metrics (Järvelin and Kekäläinen, 2002; Chapelle et al., 2009). For all aggregate metrics, we estimate uncertainty using bootstrap resampling with 10,000 samples and report 95% confidence intervals.

5 Main Results

5.1 Instruction-Following Reranking Performance

Table 2 reports aggregate performance on the 9,861-query validation benchmark. **Distilled-1B** achieves the best overall nDCG@6 among all evaluated models while remaining compact.

Three patterns stand out. First, RL-based training substantially improves over the untrained 1B backbone, raising nDCG@6 from 0.6972 to 0.7624. Second, the distilled student surpasses both the untrained 4B teacher backbone and the Stage 1 4B teacher, showing that compact rerankers can exceed larger models when distillation is performed through student-driven policy learning. Distilled-1B also exceeds the Stage 1 teacher by +7.9 points on MAIR (0.7670 vs. 0.6880); as discussed in Section 5.2, the teacher's MAIR score partly reflects specialization from off-policy RL, and we view the +4.6-point gain over offline KD as the cleaner measure of the methodological contribution. Third, Distilled-1B is competitive with or stronger than all external baselines. Distilled-1B also offers a favorable efficiency–quality tradeoff: it reduces mean latency from 27.0ms (teacher) to 9.2ms while improving ranking quality (Table 4).

5.2 Out-of-Distribution Generalization on MAIR

One feature of Table 3 deserves direct comment. Teacher GRPO reaches 0.7422 nDCG@6 on the validation benchmark but only 0.6880 on MAIR, falling below the untrained Base-1B (0.7119). We interpret this as expected: off-policy GRPO on 88K training examples specializes the teacher, reducing transfer to MAIR's argumentative, scientific, and FAQ-style subsets. The distilled student (A3) nonetheless recovers and surpasses Base-1B OOD performance despite inheriting supervision from this specialized teacher—consistent with the regularization view of on-policy distillation.

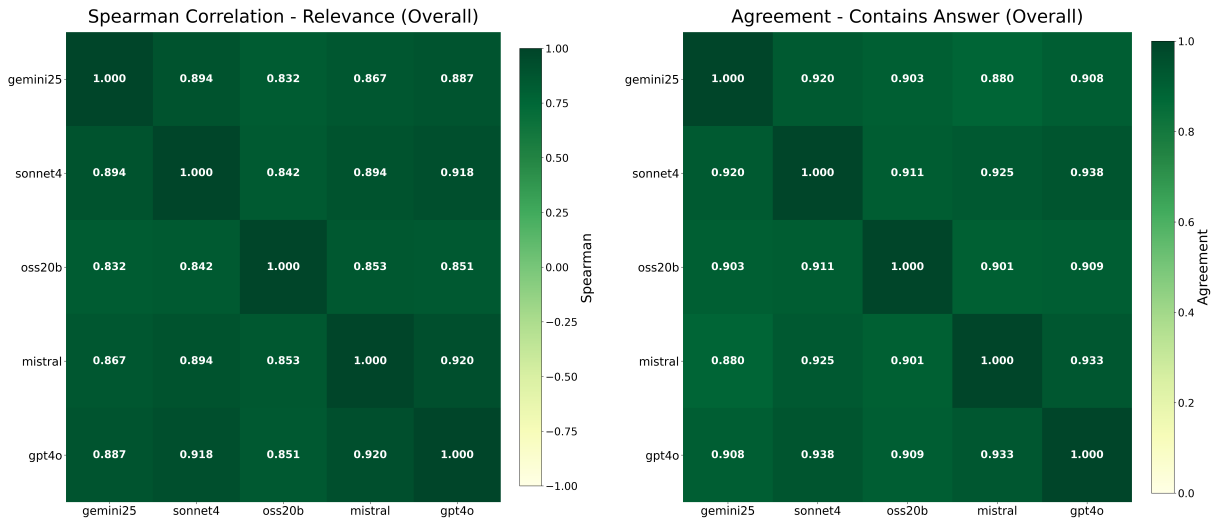
MAIR contains 869 OOD queries across 11 subsets. On-policy student learning (A3) substantially outperforms off-policy sampling (A2), offline KD (A1), and distillation from the unaligned base teacher (A8), supporting the view that robustness depends on both student-driven exploration and teacher strengthening. Hard-label training (A6) overfits in-distribution but degrades on MAIR. The advantage of on-policy distillation over offline KD is not primarily an in-distribution effect: on the validation benchmark A1 and A3 are nearly tied, while on MAIR A3 improves by +4.6 points.

6 Ablation Study

Our ablations are designed to answer three mechanistic questions, not to crown a single best configuration. (i) Does *where* supervision is evaluated matter, holding the reward source fixed? (ii) Is teacher strengthening necessary for on-policy distillation to work, or can on-policy student learning recover the same behavior from an unaligned teacher? (iii) Does the form of the reward signal (soft vs. hard) trade in-distribution fit against OOD robustness? Figure 4 summarizes the validation comparison and the validation–MAIR tradeoff across core training variants. Three findings answer these questions in turn.

Where supervision is evaluated is the key mechanism. Replacing on-policy student sampling with teacher-driven off-policy sampling (A2) consistently degrades performance, both on the validation benchmark and under distribution shift. Since A2 and A3 both use teacher-derived supervision but differ in whose policy generates the rankings, this gap isolates the role of *student exploration* rather than reward source alone. These results support our central claim that effective distillation depends on evaluating supervision on the student's own sampled ranking space.

Teacher strengthening is a necessary precondition. We additionally test whether Stage 1 teacher optimization is necessary by running on-policy GRPO from the unaligned base ZeRank-2 teacher (A8). This variant improves over offline KD and



(a) Spearman correlation: relevance

(b) Pairwise agreement: contains answer

Figure 2: Inter-judge consistency across five LLM judges used to validate the Stage 1 reward signal. Judges show strong agreement on relevance ranking and on whether a candidate passage contains the answer, which supports the use of judge-derived outputs as soft supervision for teacher training.

Model	Params	nDCG@6	SD	MRR@6	SD
Qwen3-Reranker-4B	4B	0.6564 [0.649, 0.664]	0.3821	0.6387 [0.631, 0.647]	0.4055
Base-1B (no training)	1B	0.6972 [0.690, 0.705]	0.3657	0.6639 [0.656, 0.672]	0.3909
ZeRank-2 base (4B)	4B	0.7222 [0.715, 0.730]	0.3516	0.7056 [0.698, 0.714]	0.3704
BGE Reranker v2 M3	568M	0.7310 [0.724, 0.739]	0.3697	0.7057 [0.698, 0.714]	0.3897
Teacher GRPO (4B)	4B	0.7422 [0.735, 0.750]	0.3716	0.7256 [0.718, 0.734]	0.3904
Cohere Rerank v3.5	–	0.7459 [0.7387, 0.7533]	0.3711	0.7265 [0.7190, 0.7343]	0.3889
Cohere Rerank v4.0-fast	–	0.7494 [0.7422, 0.7569]	0.3708	0.7325 [0.7251, 0.7404]	0.3875
Jina Reranker v2	278M	0.7605 [0.753, 0.768]	0.3737	0.7497 [0.742, 0.758]	0.3884
Distilled-1B (ours)	**1B**	**0.7624 [0.755, 0.770]**	**0.3730**	**0.7475 [0.740, 0.755]**	**0.3885**

Table 2: Main validation results on 9,861 instruction-following queries. Confidence intervals are bootstrap 95% CIs with 10,000 resamples.

Model	nDCG@6	SD	MRR@6	SD
Base-1B (no training)	0.7119 [0.688, 0.736]	0.3506	0.7771 [0.753, 0.802]	0.3684
Teacher GRPO (4B)	0.6880 [0.664, 0.711]	0.3508	0.7312 [0.705, 0.756]	0.3868
A1: Offline KD	0.7212 [0.716, 0.734]	0.3420	0.7860 [0.778, 0.794]	0.3720
A2: Off-policy GRPO	0.7356 [0.713, 0.759]	0.3433	0.8035 [0.779, 0.827]	0.3534
A8: On-policy GRPO from base ZeRank-2	0.7412 [0.736, 0.748]	0.3412	0.8036 [0.782, 0.832]	0.3216
A3: On-policy GRPO from Teacher (ours)	**0.7670 [0.745, 0.789]**	**0.3274**	**0.8289 [0.807, 0.851]**	**0.3323**
A4: No KL term	0.7688 [0.747, 0.790]	0.3237	0.8282 [0.806, 0.850]	0.3289
A5: No entropy	0.7665 [0.745, 0.788]	0.3285	0.8301 [0.808, 0.852]	0.3328
A6: Hard labels	0.7365 [0.713, 0.759]	0.3455	0.8000 [0.777, 0.823]	0.3512

Table 3: OOD results on MAIR. Full ablation table in Appendix C.

the off-policy student, but remains well below A3 on MAIR. On the validation benchmark, A8 remains competitive on several subsets, but its main value in the paper is as an OOD ablation showing that teacher strengthening improves the effectiveness of subsequent on-policy distillation. The result indicates that the gains of the two-stage pipeline do not come from student-side on-policy learning alone: they also depend on first improving the teacher with instruction-aware RL before distillation.

Soft rewards trade in-distribution fit for OOD robustness. Replacing soft teacher rewards with hard labels (A6) yields the best validation performance among the core variants, but substantially reduces OOD performance on MAIR. This contrast exposes a clear accuracy–generalization tradeoff: hard labels fit the in-distribution benchmark more aggressively, whereas soft rewards preserve relative

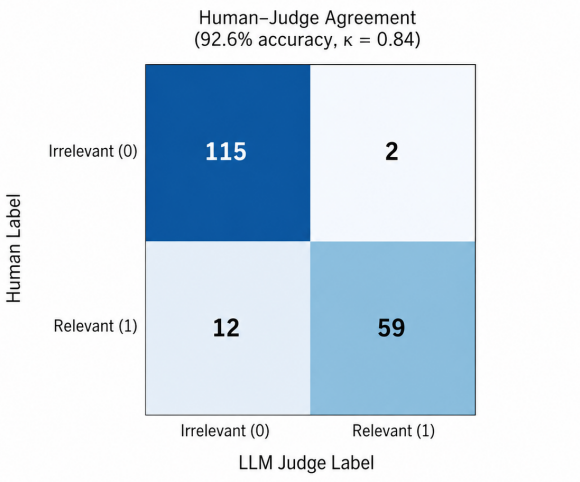


Figure 3: Agreement between the Stage 1 LLM judge and human relevance annotations. The judge achieves 92.6% agreement and Cohen's $\kappa = 0.84$, with most disagreements coming from false negatives, indicating a conservative bias relative to human labeling.

preference information that regularizes the student under distribution shift. For deployment-oriented instruction-following reranking, this tradeoff favors soft supervision over maximal validation fit.

Policy-gradient training is the dominant learning signal. Removing the KL term (A4) or entropy bonus (A5) leaves performance close to the full objective, indicating that policy-gradient optimization with teacher rewards is the primary source of improvement. A5 is numerically very close to A3 on both validation and MAIR, while A4 is slightly stronger than A3 both in aggregate and in paired query-level tests. We therefore do not claim that the full objective is uniquely best-performing in every setting. Instead, we view A3 as a robust reference configuration that preserves the complete exploration-aware formulation of the method, while A4 and A5 show that the auxiliary KL and entropy terms are secondary to the main gain from student-driven on-policy distillation with soft teacher rewards.

To complement the aggregate tables, we also compare key variants with paired Wilcoxon signed-rank tests over query-level scores. The proposed on-policy student (A3) significantly outperforms offline KD (A1) on MAIR (two-sided $p = 1.95 \times 10^{-9}$; one-sided A3>A1 $p = 9.76 \times 10^{-10}$), strengthening the claim that on-policy distillation yields substantially more robust ranking behavior under distribution shift. In contrast, the no-KL variant (A4) achieves slightly stronger query-level performance than A3 on MAIR (two-sided $p = 0.0269$; one-sided A3<A4 $p = 0.0134$). However, this aggregate result does not imply uniform dominance: the subset-level picture remains mixed, with A3 outperforming A4 on some MAIR subsets and A4 stronger on others. We therefore interpret the core mechanism of the paper as student-driven on-policy distillation with soft teacher rewards, while treating the KL term as an optional regularizer rather than a uniformly beneficial component.

To further illustrate out-of-distribution behavior, Figure 5 shows MAIR performance across the core training variants and representative reference models. The same pattern observed in Table 3 holds visually at both nDCG@6 and MRR@6: on-policy student-driven distillation (A3) substantially improves over the teacher and the off-policy variant (A2), while hard-label training (A6) underperforms A3 under distribution shift despite its stronger in-distribution fit. The figure also makes clear that the gains are not limited to a single metric, but are consistent across both ranking quality and first-hit retrieval quality.

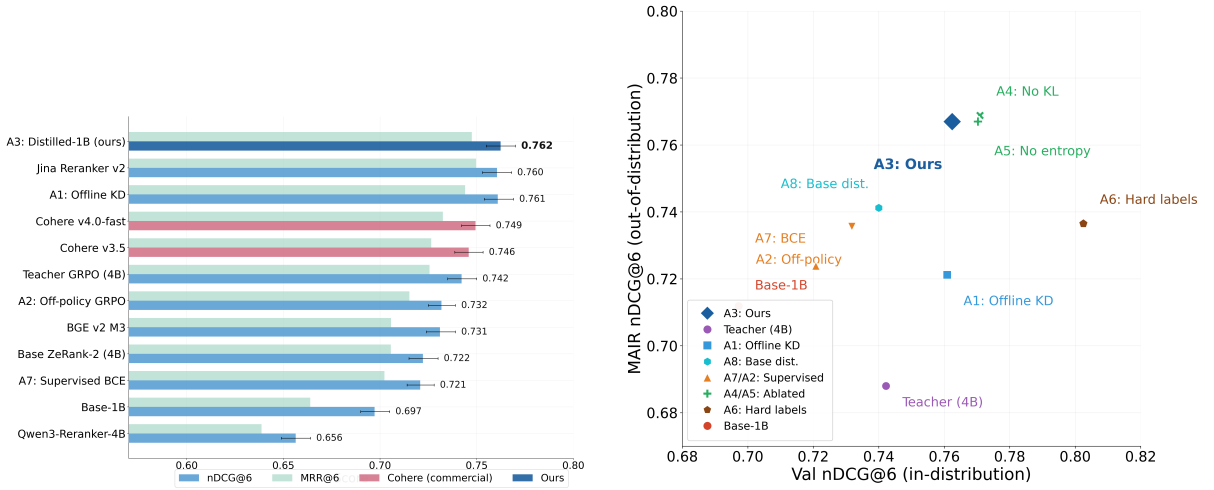
6.1 Hyperparameter Sensitivity

We next test the robustness of the default training configuration by sweeping group size, Plackett–Luce temperature, KL weight, and entropy coefficient. Across these sweeps, validation performance remains nearly flat, indicating that the method is not highly sensitive to hyperparameter tuning.

The no-entropy setting appears both as the Set A core ablation A5 and as the Set B sweep point B4 ($\text{ent} = 0.0$); the two are numerically aligned as expected. MAIR shows somewhat stronger sensitivity than the validation benchmark, with the clearest variation appearing in the KL sweep: stronger KL regularization improves OOD performance while leaving validation nearly unchanged.

This stability is important for two reasons. First, it shows that the observed gains are not the result of narrow tuning around a single configuration. Second, it clarifies the role of the auxiliary terms: while validation performance changes little across the sweeps, MAIR shows modest sensitivity to the KL coefficient, with stronger KL regularization improving robustness under distribution shift. This is consistent with the view that KL regularization is not the main driver of learning, but can still improve generalization stability. The full Set B plot and numeric tables are provided in Appendix D.

Overall, the ablations support a simple interpre-



(a) Validation comparison across major baselines and ablations.

(b) Validation–MAIR tradeoff across core training variants. Top-right is better.

Figure 4: Core ablation evidence for the proposed training pipeline. On-policy student-driven distillation provides the strongest overall tradeoff between in-distribution fit and OOD generalization, while hard-label supervision improves validation performance but harms transfer under distribution shift.

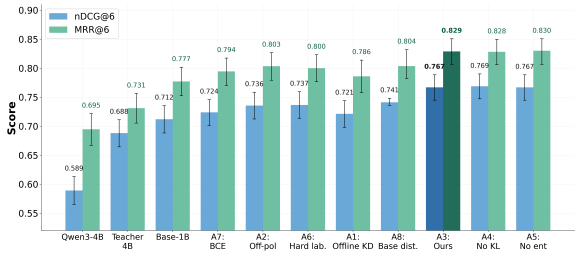


Figure 5: MAIR out-of-distribution performance across representative reference models and core training variants. On-policy distillation (A3) substantially improves over the teacher and off-policy training, while hard-label supervision (A6) reduces robustness relative to soft teacher rewards. Error bars indicate bootstrap 95% confidence intervals.

Model	Params	p50	p95	Mean±SD
Distilled-1B (ours)	1235.8M	9.2	10.0	9.2 ± 0.3
ZeRank-2 Teacher (4B)	4022.5M	24.9	41.8	27.0 ± 4.0
BGE Reranker v2 M3	567.8M	7.9	16.5	9.2 ± 2.5
Jina Reranker v2	278.4M	5.5	18.7	7.6 ± 3.5
Qwen3-Reranker-4B	4021.8M	37.9	55.0	38.9 ± 4.5

Table 4: Single-query latency under a fixed hardware setup.

tation of the method: the primary benefit comes from *on-policy policy-gradient training with soft teacher rewards*, while KL and entropy terms serve as secondary regularizers rather than essential components.

7 Efficiency Analysis

A central motivation for compact rerankers is deployment efficiency. Table 4 compares single-query inference latency under a fixed hardware setup.

Distilled-1B offers a favorable efficiency–quality tradeoff. It is substantially faster than 4B rerankers while remaining competitive with or stronger than all evaluated baselines on ranking quality. Compared to the teacher, the student reduces mean latency from 27.0ms to 9.2ms while also improving validation performance.

8 Discussion and Conclusion

Our results support three conclusions. First, the main gain comes from *where* supervision is evaluated: the A2–A3 gap shows that evaluating teacher-derived supervision on the student's own sampled rankings improves both in-distribution performance and OOD generalization.

Second, teacher strengthening is necessary but not sufficient. A8 improves over offline KD and off-policy student training, but remains well below the full pipeline on MAIR, showing that Stage 1 and Stage 2 are complementary.

Third, soft rewards improve robustness under shift. A6 achieves the strongest validation fit but degrades on MAIR, indicating that soft teacher rewards act as an implicit regularizer. The closely related variants A3, A4, and A5 remain tightly clustered, suggesting that the main mechanism is student-driven on-policy learning with soft teacher rewards, while KL and entropy are optional stabilizers.

9 Limitations

Our study focuses on text-only reranking and English-language instruction following. The teacher is trained with judge-derived reward signals, so the method inherits any bias or calibration error in the judge. Our MAIR evaluation uses subsets topically relevant to our downstream enterprise setting rather than the full benchmark, which narrows the OOD claim. Most ablations use a single random seed, so small differences among closely clustered variants (A3, A4, A5) should be interpreted with caution (see Appendix E.3 for multi-seed results). Finally, the extent to which the student-surpasses-teacher finding transfers across architectures, domains, and teacher strengths remains an open question.

10 Ethical Considerations

This work studies instruction-following reranking for text retrieval and does not introduce new user-facing generation capabilities. The main ethical risks arise from misranking rather than content generation: instruction-following rerankers may amplify bias present in training data, judge models, or candidate corpora, and may under-rank relevant documents for particular domains, perspectives, or user groups. Because our Stage 1 teacher is trained with judge-derived reward signals, the pipeline can also inherit calibration errors or systematic preferences from the judge model. We partially mitigate this risk by validating the judge against both other strong LLM evaluators and human relevance annotations, but these checks do not eliminate all bias.

Our experiments are limited to English text-only reranking and to the datasets and MAIR subsets described in the paper, so the findings should not be interpreted as guaranteeing fair or robust behavior across all domains or deployment settings. In enterprise use, instruction-following rerankers should be monitored with domain-appropriate evaluation, auditing, and human oversight, especially in high-stakes settings such as legal, medical, financial, or employment-related retrieval.

References

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. 2024. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations (ICLR)*.

Akari Asai, Timo Schick, Patrick Lewis, Xilun Chen, Gautier Izacard, Sebastian Riedel, Hannaneh Hajishirzi, and Wen-tau Yih. 2023. Task-aware retrieval with instructions. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3650–3675.

Olivier Chapelle, Donald Metlzer, Ya Zhang, and Pierre Grinspan. 2009. Expected reciprocal rank for graded relevance. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, pages 621–630.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. 2024. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*.

Tommaso Furlanello, Zachary C. Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. 2018. Born again neural networks. *arXiv preprint arXiv:1805.04770*.

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

Sebastian Hofstätter, Sheng-Chieh Lin, Jheng-Hong Yang, Jimmy Lin, and Allan Hanbury. 2021. Efficiently teaching an effective dense retriever with balanced topic aware sampling. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 113–122.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems*, 20(4):422–446.

Omar Khattab and Matei Zaharia. 2020. ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. *arXiv preprint arXiv:2004.12832*.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. 2020. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*.

Kevin Lu and Thinking Machines Lab. 2025. On-policy distillation. https://thinkingmachines.ai/blog/on-policy-distillation/. Thinking Machines Lab blog post.

Xueguang Ma, Xinyu Zhang, Ronak Pradeep, and Jimmy Lin. 2023. Zero-shot listwise document reranking with a large language model. *arXiv preprint arXiv:2305.02156*.

Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with BERT. *arXiv preprint arXiv:1901.04085*.

NVIDIA. 2025. nvidia/llama-nemotron-rerank-1b-v2. Hugging Face model card. Accessed 2026-04-10.

Hanseok Oh, Hyunji Lee, Seonghyeon Ye, Haebin Shin, Hansol Jang, Changwook Jun, and Minjoon Seo. 2024. INSTRUCTIR: A benchmark for instruction following of information retrieval models. *arXiv preprint arXiv:2402.14334*.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

Ronak Pradeep, Sahel Sharifymoghaddam, and Jimmy Lin. 2023a. Rankvicuna: Zero-shot listwise document reranking with open-source large language models. *arXiv preprint arXiv:2309.15088*.

Ronak Pradeep, Sahel Sharifymoghaddam, and Jimmy Lin. 2023b. Rankzephyr: Effective and robust zero-shot listwise reranking is a breeze! *arXiv preprint arXiv:2312.02724*.

Zhen Qin, Rolf Jagerman, Kai Hui, Honglei Zhuang, Junru Wu, Le Yan, Jiaming Shen, Tianqi Liu, Jialu Liu, Donald Metzler, Xuanhui Wang, and Michael Bendersky. 2024. Large language models are effective text rankers with pairwise ranking prompting. In *Findings of the Association for Computational Linguistics: NAACL 2024*.

Qwen Team. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

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. *arXiv preprint arXiv:2305.18290*.

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*.

Orion Weller, Benjamin Chang, Sean MacAvaney, Kyle Lo, Arman Cohan, Benjamin Van Durme, Dawn Lawrie, and Luca Soldaini. 2024a. FollowIR: Evaluating and teaching information retrieval models to follow instructions. *arXiv preprint arXiv:2403.15246*.

Orion Weller, Benjamin Van Durme, Dawn Lawrie, Ashwin Paranjape, Yuhao Zhang, and Jack Hessel. 2024b. Promptriever: Instruction-trained retrievers can be prompted like language models. *arXiv preprint arXiv:2409.11136*.

Andrew Yates, Rodrigo Nogueira, and Jimmy Lin. 2021. Pretrained transformers for text ranking: BERT and beyond. In *Proceedings of NAACL-HLT 2021: Tutorials*, pages 1–4.

ZeroEntropy. 2025. zeroentropy/zerank-2-reranker. https://huggingface.co/zeroentropy/zerank-2-reranker. Hugging Face model card, 4B instruction-following reranker.

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. *arXiv preprint arXiv:2306.05685*.

A Reproducibility and Compute

To support reproducibility, we provide anonymous artifacts for the final Distilled-1B checkpoint, evaluation datasets, and evaluation code. All repositories listed below are publicly accessible and require no authentication to download.

- **Model checkpoint:** https://huggingface.co/anonymousauthor01/instruction_following_reranker
- **Validation dataset:** https://huggingface.co/datasets/anonymousauthor01/emnlp-2026-ifr-train-val-set
- **MAIR OOD subsets:** https://huggingface.co/datasets/anonymousauthor01/emnlp-2026-ifr-mair-ood
- **Code repository with instructions:** https://github.com/anonymousauthor01-conf/emnlp-2026-artifact-release

The code repository contains the evaluation scripts, one-command evaluation launcher, dependency specification, and instructions for automatically downloading the checkpoint and datasets from HuggingFace. The released evaluation package reproduces the reported validation and MAIR results for the final model using the same candidate sets, relevance labels, metrics, and bootstrap confidence-interval procedure used in the paper. Evaluation scripts and configurations for the ablation variants (A1–A8, Set B hyperparameter sweeps) will be released upon acceptance alongside the full training pipeline.

Upon acceptance, we will release the full training scripts, data preprocessing pipeline, LLM-judge prompt templates, the exact configuration files used for the reported runs, and the scripts and configurations used to reproduce the ablation studies.

All reported training and evaluation runs were conducted on NVIDIA H200 GPUs. The Stage 1 teacher run required approximately 3 days and 23 hours on 2 H200 GPUs, the default Stage 2 on-policy distillation run required approximately 4 hours and 50 minutes on a single H200, and the full validation+MAIR evaluation suite required approximately 20 minutes on a single GPU. The core reported pipeline therefore required approximately 195 H200 GPU-hours in total, excluding non-reported debugging and exploratory runs.

B Additional Experimental Details

This appendix provides supplementary evidence for the main empirical claims of the paper. In the main text, we focus on the central results: (i) on-policy student exploration is the primary driver of effective distillation, (ii) soft teacher rewards improve robustness under distribution shift, and (iii) the resulting 1B model offers a strong quality–efficiency tradeoff. Here we include the full MAIR ablation table, the full Set B numeric ablation tables, a focused per-subset comparison of A3/A4/A5, and expanded validation-set analyses comparing all major baselines at the subset level.

A recurring pattern across these appendix results is that the same training choice that is strongest in the main paper—student-driven on-policy distillation with soft teacher rewards—also remains the most stable choice under shift and across datasets. At the same time, the appendix helps qualify the main story: closely related variants such as A4 and A5 remain very close to A3, and can outperform it on individual subsets, but these differences are small compared with the larger gap between on-policy student-driven distillation and the offline or off-policy alternatives.

C Full MAIR Ablation Results

Table 5 reports the complete MAIR ablation results for all core Set A variants. The compact MAIR table in the main paper includes only the most diagnostic rows; this full version shows the broader ordering across baselines and ablations. Two patterns are worth highlighting. First, on-policy student distillation remains substantially stronger than both the teacher and offline KD under distribution shift. A8 further shows that teacher quality matters as well: running on-policy GRPO from the unaligned base ZeRank-2 teacher remains clearly below A3, indicating that the two-stage pipeline benefits not only from on-policy student learning but also from first strengthening the teacher with instruction-aware RL. Second, the hard-label variant achieves strong in-distribution performance but drops noticeably on MAIR, reinforcing the interpretation that soft rewards act as a regularizer for OOD robustness.

C.1 Why does Teacher GRPO underperform Base-1B on MAIR?

The aggregate MAIR results might suggest that the Stage 1 teacher (**Teacher GRPO**) simply trans-

Model	nDCG@6	SD	95% CI	MRR@6	SD
Base-1B (no training)	0.7119	0.3506	[0.688, 0.736]	0.7771	0.3684
ZeRank-2 base (4B)	0.6938	0.3482	[0.670, 0.717]	0.7266	0.3850
Teacher GRPO (4B)	0.6880	0.3508	[0.664, 0.711]	0.7312	0.3868
A7: Supervised BCE	0.7239	0.3364	[0.701, 0.746]	0.7945	0.3505
A2: Off-policy GRPO	0.7356	0.3433	[0.713, 0.759]	0.8035	0.3534
A8: On-policy GRPO from base ZeRank-2	0.7412	0.3412	[0.736, 0.748]	0.8036	0.3216
A1: Offline KD	0.7212	0.3420	[0.716, 0.734]	0.7860	0.3720
A3: On-policy GRPO from Teacher (ours)	**0.7670**	**0.3274**	**[0.745, 0.789]**	**0.8289**	**0.3323**
A4: No KL term	0.7688	0.3237	[0.747, 0.790]	0.8282	0.3289
A5: No entropy	0.7665	0.3285	[0.745, 0.788]	0.8301	0.3328
A6: Hard labels	0.7365	0.3455	[0.713, 0.759]	0.8000	0.3512

Table 5: Full MAIR ablation results. Confidence intervals are bootstrap 95% CIs with 10,000 resamples.

fers worse than the untrained **Base-1B** model under distribution shift. A subset-level analysis shows a more specific pattern. Teacher GRPO does *not* degrade uniformly across MAIR: relative to Base-1B, it is stronger on 7 of 11 subsets, including ArguAna, DD_2016, FiQA, SciDocs, NFCorpus, Quora, and Trec-Covid. However, this improvement is offset by a small number of large OOD failures, most notably on LitSearch and Touche, which dominate the aggregate average.

This pattern suggests that Stage 1 off-policy RL does not merely make the teacher weaker OOD; rather, it produces a more *specialized* ranking policy. The teacher remains strong on several subsets, but becomes brittle on a few domains that are farther from the effective support of the instruction-following training distribution. In particular, the extremely large drop on LitSearch indicates that the teacher's OOD weakness is concentrated rather than uniform.

The final A3 student does not uniformly dominate the teacher either. Instead, its main gain comes from *repairing the teacher's largest OOD failures*. A3 recovers dramatically on LitSearch, Touche, and SciFact, while remaining slightly below the teacher on a number of subsets such as ArguAna, DD_2016, SciDocs, and Trec-Covid. This is consistent with the paper's main interpretation: on-policy student distillation does not simply inherit the Stage 1 teacher, but acts as a corrective mechanism by evaluating supervision on rankings sampled from the student's own policy.

D Full Set B Numeric Ablation Results

The main paper summarizes Set B qualitatively through the robustness discussion. Here we provide the full plot and numeric tables for validation and MAIR so that the stability claims can be inspected directly. These results confirm that the method is not highly sensitive to the exact default configuration: validation performance remains nearly flat across all sweeps, while MAIR shows only modest variation, most visibly in the KL sweep.

D.1 Validation Results

Table 7 confirms that the in-distribution benchmark is highly stable with respect to group size, sampling temperature, KL weight, and entropy coefficient. The narrow spread of scores suggests that the main gains are not coming from brittle tuning around a single configuration.

D.2 MAIR Results

Table 8 shows the same sweeps on MAIR. While the overall method remains stable, MAIR displays somewhat stronger sensitivity to the KL coefficient than the validation set. In particular, larger KL values improve OOD performance, consistent with our interpretation in the main text that KL is not the main source of learning but can still improve generalization stability.

E Per-Dataset Validation Breakdown Across All Baselines

Aggregate validation results in the main paper hide meaningful differences across datasets. Table 9 provides the full per-dataset validation breakdown for all major baselines and the final distilled model. This view makes two points clear. First, the distilled student is not winning only because of a single dataset; it remains competitive across web search, code, math, instruction-following, and multi-hop settings. Second, different baselines exhibit different strengths: for example, ZeRank-2 base is already strong on FollowIR, InstructIR, and several InfIR subsets, Cohere v4.0-fast is particularly strong on Robust04 and LeetCode, and Jina v2 is strong on MS MARCO and Robust04.

MAIR subset	Base-1B	Teacher GRPO	A3	Teacher–Base	A3–Teacher
ArguAna	0.374	0.600	0.462	+0.227	-0.138
Core_2017	0.656	0.629	0.675	-0.027	+0.046
DD_2016	0.530	0.701	0.527	+0.171	-0.175
FiQA	0.832	0.931	0.965	+0.099	+0.034
LitSearch	0.982	0.125	0.995	-0.858	+0.870
NFCorpus	0.635	0.669	0.649	+0.034	-0.021
Quora	0.951	1.000	0.978	+0.049	-0.022
SciDocs	0.420	0.651	0.473	+0.231	-0.178
SciFact	0.705	0.700	0.835	-0.004	+0.134
Trec-Covid	0.870	0.915	0.897	+0.045	-0.018
Touche	0.930	0.751	0.941	-0.179	+0.190

Table 6: Subset-level MAIR diagnosis of Base-1B, Teacher GRPO, and A3 using nDCG@6. Teacher GRPO is stronger than Base-1B on most subsets, but suffers a few large OOD failures, especially on LitSearch and Touche, which dominate its aggregate MAIR drop. A3 does not uniformly beat the teacher; instead, its overall gain comes primarily from correcting those largest OOD failures.

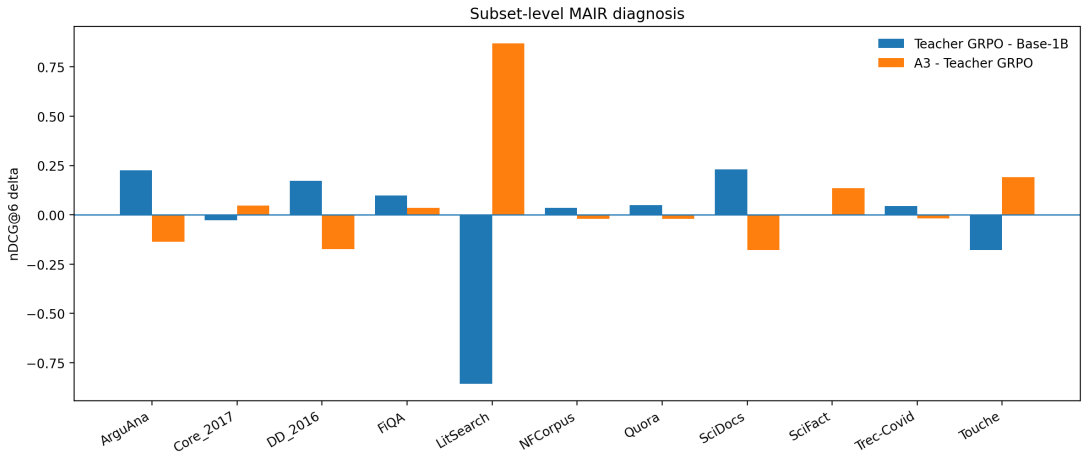


Figure 6: Subset-level MAIR diagnosis using nDCG@6 deltas. The left bar for each subset shows the change from Base-1B to Teacher GRPO; the right bar shows the change from Teacher GRPO to A3. Teacher GRPO improves over Base-1B on most subsets, but suffers a few concentrated OOD failures, especially on LitSearch and Touche. A3's aggregate gain comes mainly from repairing those largest teacher failures rather than uniformly dominating the teacher on every subset.

Variant	nDCG@6	SD	95% CI	MRR@6	SD
B1: group_size=4	0.7607	0.3729	[0.753, 0.768]	0.7454	0.3887
B1: group_size=6	0.7611	0.3730	[0.754, 0.769]	0.7458	0.3887
B1: group_size=12	0.7630	0.3731	[0.756, 0.771]	0.7484	0.3886
B2: temp=0.5	0.7616	0.3729	[0.754, 0.769]	0.7465	0.3886
B2: temp=0.7	0.7611	0.3729	[0.754, 0.769]	0.7459	0.3885
B2: temp=1.5	0.7625	0.3730	[0.755, 0.770]	0.7478	0.3886
B3: kl=0.1	0.7625	0.3731	[0.755, 0.770]	0.7476	0.3885
B3: kl=0.5	0.7625	0.3731	[0.755, 0.770]	0.7476	0.3886
B3: kl=2.0	0.7618	0.3731	[0.755, 0.769]	0.7473	0.3886
B4: ent=0.0	0.7624	0.3731	[0.755, 0.770]	0.7476	0.3885
B4: ent=0.05	0.7620	0.3730	[0.755, 0.770]	0.7470	0.3886
B4: ent=0.1	0.7613	0.3730	[0.754, 0.769]	0.7460	0.3886

Table 7: Full Set B validation results. In-distribution performance remains highly stable across all sweeps.

At the same time, the distilled student does not dominate every validation subset. In particular, some baselines remain clearly stronger on Robust04, and InfoSearch is more tightly clustered across models. We therefore interpret the method as improving the overall quality–robustness trade-off rather than as establishing uniform superiority on every in-distribution subset.

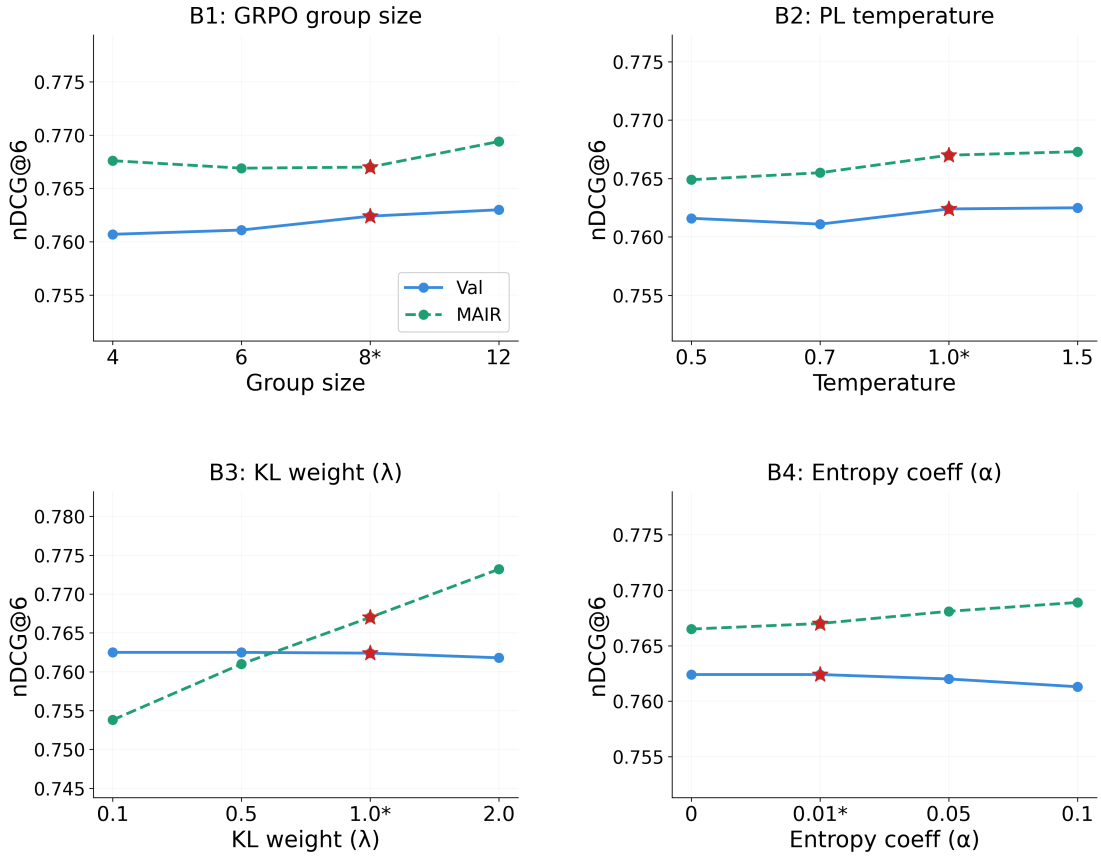


Figure 7: Hyperparameter sensitivity of the default training setup. Validation performance is nearly flat across all sweeps, while MAIR shows modest sensitivity primarily to the KL weight. Red stars mark the default configuration used in the main results.

Variant	nDCG@6	SD	95% CI	MRR@6	SD
B1: group_size=4	0.7676	0.3275	[0.745, 0.789]	0.8319	0.3318
B1: group_size=6	0.7669	0.3279	[0.745, 0.789]	0.8305	0.3330
B1: group_size=12	0.7694	0.3261	[0.748, 0.791]	0.8328	0.3309
B2: temp=0.5	0.7649	0.3267	[0.743, 0.787]	0.8306	0.3313
B2: temp=0.7	0.7655	0.3268	[0.744, 0.788]	0.8305	0.3315
B2: temp=1.5	0.7673	0.3283	[0.745, 0.789]	0.8314	0.3327
B3: kl=0.1	0.7538	0.3360	[0.732, 0.776]	0.8142	0.3418
B3: kl=0.5	0.7610	0.3297	[0.739, 0.783]	0.8243	0.3365
B3: kl=2.0	0.7732	0.3210	[0.752, 0.795]	0.8378	0.3258
B4: ent=0.0	0.7665	0.3285	[0.745, 0.788]	0.8301	0.3328
B4: ent=0.05	0.7681	0.3262	[0.746, 0.790]	0.8316	0.3311
B4: ent=0.1	0.7689	0.3251	[0.747, 0.791]	0.8326	0.3298

Table 8: Full Set B MAIR results. OOD performance is broadly stable, with the most visible sensitivity appearing in the KL sweep.

Dataset	Qwen3-4B	Base-1B	ZeRank-2 base	BGE v2 M3	Teacher GRPO	Cohere v3.5	Cohere v4-fast	Jina v2	A3 D
FollowIR	0.910	0.947	1.000	0.962	1.000	0.970	1.000	0.993	
InstructIR	0.760	0.956	1.000	1.000	1.000	1.000	1.000	1.000	
InfoSearch	0.829	0.835	0.831	0.831	0.843	0.848	0.844	0.843	
MS MARCO	0.916	0.811	0.892	0.847	0.897	0.879	0.874	0.943	
MetaMath	0.728	0.797	0.926	0.950	0.926	0.948	0.971	0.822	
LeetCode	0.717	0.745	0.712	0.733	0.724	0.778	0.831	0.754	
Robust04	0.952	0.605	0.605	0.578	0.620	0.660	0.819	0.970	
WebQA	0.255	0.444	0.422	0.458	0.435	0.455	0.453	0.435	

Table 9: Per-dataset validation nDCG@6 across all major baselines and the final distilled model.

E.1 Per-Subset Comparison of A3, A4, and A5

To better understand the role of the KL and entropy terms, we compare A3 (full objective), A4 (without KL), and A5 (without entropy) at the per-subset level on MAIR. The three variants are very close in aggregate, and the subset-level picture remains mixed: wins are split across domains, and none of the three variants consistently dominates the others. This supports the interpretation in the main paper that policy-gradient training with teacher-derived soft rewards is the primary source of improvement, while KL and entropy act as auxiliary stabilizers rather than indispensable components.

A4 is slightly strongest overall on MAIR, while A3 and A5 remain very close to one another. A3 remains strongest on several OOD subsets, including Core_2017, DD_2016, Trec-Covid, and LitSearch, while A4 is strongest on others, including ArguAna, FiQA, Quora, and Touche. A5 is also competitive across most subsets. Overall, the absence of a consistent winner suggests that the differences among A3, A4, and A5 are modest relative to the larger gap between on-policy student-driven distillation and the offline or off-policy alternatives.

E.2 Paired Statistical Tests for Key Variant Comparisons

To complement the aggregate bootstrap confidence intervals reported in the main paper, we additionally compare key variants using paired Wilcoxon signed-rank tests over query-level nDCG@6 scores. This analysis focuses on the two most important comparisons for the paper's claims: A3 versus A1, which tests whether on-policy distillation improves over offline KD, and A3 versus A4, which tests whether the KL term materially changes the behavior of the full objective.

Table 11 shows two qualitatively different outcomes. First, A3 strongly and significantly outperforms A1 on both validation and MAIR, which reinforces the main claim that on-policy student-driven distillation is substantially stronger than offline KD. Second, A4 is slightly but significantly stronger than A3 in paired tests, especially on validation. At the same time, the per-subset MAIR results do not show uniform dominance: A3 is strongest on several subsets, while A4 is strongest on others. We therefore interpret A3 as a useful reference formulation of the full objective rather than as the uniquely best-performing member of this closely related family.

Figure 8 provides a compact visual summary of these paired effect estimates. The plot makes clear that the A3–A1 advantage is large and consistently positive on both validation and MAIR, whereas the A3–A4 difference is much smaller in magnitude and favors A4 in aggregate.

E.3 Multi-seed stability of A3, A4, and A5

A possible concern in Section 6 is that the close ordering among A3, A4, and A5 could reflect single-seed noise rather than a stable relationship between these objective variants. We therefore reran the tightly clustered A3/A4/A5 family with 3 random seeds and evaluated each run on both the validation benchmark and MAIR.

We focus on A3, A4, and A5 because these are the variants whose single-seed results are closest in both validation and OOD evaluation, and hence are the only ones for which training-seed variance could materially affect the interpretation. In contrast, the larger gaps for A1, A2, A6, and A8 are already well separated from A3 in the main results and are less likely to be explained by small seed-level fluctuations.

Table 12 reports the aggregate 3-seed results, and Figure 9 visualizes the same comparison for validation and MAIR. The multi-seed results confirm that A3, A4, and A5 remain tightly clustered across seeds. A4 is marginally strongest on the validation benchmark, while A3 is marginally strongest on MAIR, with A5 remaining close to both. This reinforces our interpretation that the main gain comes from student-driven on-policy learning with soft teacher rewards, while the KL and entropy terms act as secondary stabilizers rather than determining the core improvement. We therefore treat A3/A4/A5 as a closely related family of strong variants rather than claiming a single uniformly dominant configuration.

E.4 Qualitative Comparisons

Table 13 shows four representative qualitative comparisons centered on the final A3 student. We include one example each for offline KD (A1), the Stage 1 teacher, hard-label training (A6), and off-policy student distillation (A2). In all four cases, A3 flips the rank-1 decision: the relevant document is promoted to the top position, whereas the comparison model assigns rank 1 to a near-miss or distractor. These examples are intended as illustrations of the aggregate trends in the main paper rather than as standalone evidence.

MAIR subset	nDCG@6			MRR@6		
	A3	A4	A5	A3	A4	A5
ArguAna	0.4625	**0.5444**	0.4670	0.4147	**0.4975**	0.4178
Core_2017	**0.6745**	0.6720	0.6738	**1.0000**	**1.0000**	**1.0000**
DD_2016	**0.5267**	0.5171	0.5147	**0.8450**	0.8317	0.8367
FiQA	0.9646	**0.9687**	0.9645	0.9833	**0.9867**	0.9833
LitSearch	**0.9950**	0.9943	0.9943	**0.9933**	0.9900	0.9925
NFCorpus	0.6487	0.6334	**0.6505**	0.7522	0.7278	**0.7588**
Quora	0.9784	**0.9788**	0.9788	0.9728	**0.9733**	**0.9733**
SciDocs	0.4729	0.4355	**0.4755**	0.6640	0.6233	**0.6717**
SciFact	**0.8346**	0.8239	0.8303	**0.8103**	0.8033	0.8078
Trec-Covid	**0.8970**	0.8872	0.8922	**1.0000**	**1.0000**	**1.0000**
Touche	0.9415	**0.9493**	0.9431	**1.0000**	**1.0000**	**1.0000**
Overall MAIR	0.7670	**0.7688**	0.7665	0.8289	0.8282	**0.8301**

Table 10: Per-subset MAIR comparison for A3 (full objective), A4 (no KL), and A5 (no entropy), reported for both nDCG@6 and MRR@6. All three variants are very close in aggregate, and wins remain split across subsets rather than dominated by a single configuration.

Comparison	Split	Mean diff.	95% CI	Wilcoxon p (2-sided)	Direction
A3 vs A1	Validation	+0.0643	[0.0613, 0.0673]	0.0	A3 > A1
A3 vs A1	MAIR	+0.0220	[0.0151, 0.0291]	1.95×10^{-9}	A3 > A1
A3 vs A4	Validation	-0.0084	[-0.0098, -0.0070]	5.91×10^{-33}	A3 < A4
A3 vs A4	MAIR	-0.0043	[-0.0076, -0.0012]	2.69×10^{-2}	A3 < A4

Table 11: Paired Wilcoxon signed-rank tests over query-level nDCG@6 scores for the two most important variant comparisons. Positive mean differences favor A3; negative mean differences favor the comparison variant.

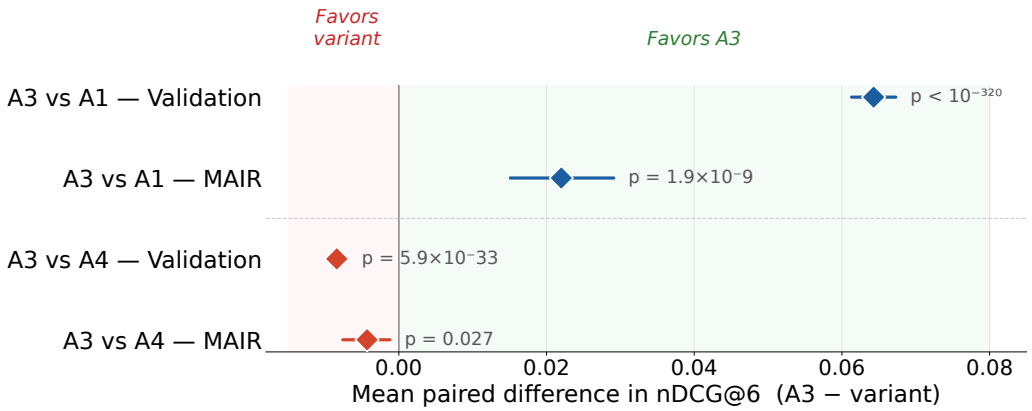


Figure 8: Forest plot of paired query-level nDCG@6 effect estimates for the two key variant comparisons. Positive values favor A3; negative values favor the comparison variant. A3 shows a strong positive advantage over offline KD (A1) on both validation and MAIR, while A4 is slightly stronger than A3 in aggregate paired comparisons. Horizontal bars show bootstrap 95% confidence intervals.

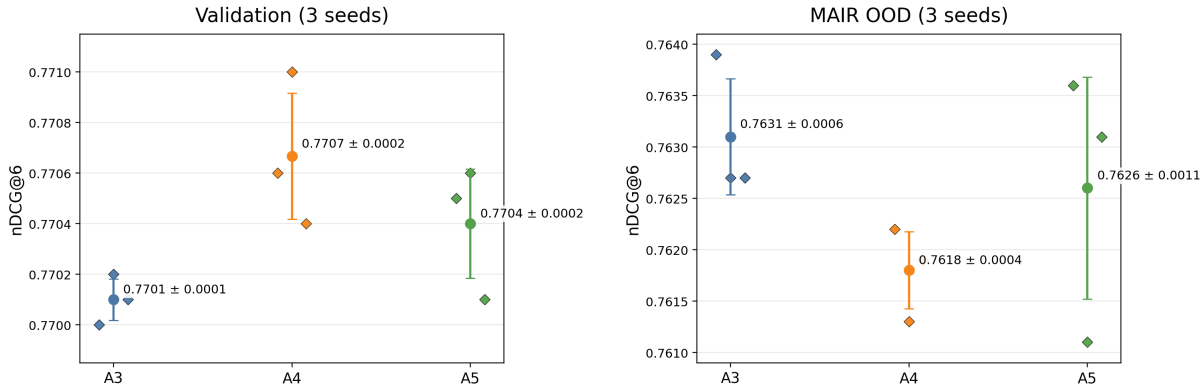
Two patterns recur across the examples. First, many failures are not gross topical errors but *ranking* errors: the competing model retrieves a plausible candidate set but places the best instruction- or claim-aligned passage below less relevant distractors. Second, the A3 student often improves by making finer distinctions among highly similar candidates rather than by discovering an entirely different document. This is consistent with the broader empirical picture in the paper: student-driven on-policy distillation improves the ordering of candidate documents even when the relevant passage is already present in the pool.

E.5 Offline KD vs On-Policy Distillation

A1 and A3 are intentionally close on the in-distribution validation benchmark: A1 reaches 0.7609 nDCG@6 and A3 reaches 0.7624, with heavily overlapping confidence intervals. We therefore do not interpret the benefit of on-policy distilla-

Variant	# Seeds	Val nDCG@6	Val MRR@6	MAIR nDCG@6	MAIR MRR@6
A3	3	0.7701 ± 0.0001	0.7449 ± 0.0002	0.7631 ± 0.0007	0.8252 ± 0.0006
A4	3	0.7707 ± 0.0003	0.7490 ± 0.0004	0.7618 ± 0.0005	0.8229 ± 0.0011
A5	3	0.7704 ± 0.0003	0.7486 ± 0.0004	0.7626 ± 0.0013	0.8247 ± 0.0022

Table 12: Aggregate multi-seed results for the tightly clustered A3/A4/A5 variants. Values are mean ± standard deviation across 3 random seeds.



(a) Validation benchmark over 3 random seeds.

(b) MAIR OOD benchmark over 3 random seeds.

Figure 9: Multi-seed stability of the tightly clustered A3/A4/A5 family over 3 random seeds. Dots show individual seeds; circles and error bars show mean ± standard deviation across seeds. A4 is marginally strongest on the validation benchmark, while A3 is marginally strongest on MAIR. The small spread across seeds supports our interpretation that these three variants are closely related, and that the main gain comes from student-driven on-policy learning with soft teacher rewards, while KL and entropy act as secondary stabilizers.

tion as a stronger in-distribution fitting mechanism than offline KD. Instead, the distinction becomes visible under distribution shift. On MAIR, A3 improves over A1 from 0.7414 to 0.7670 nDCG@6 and from 0.8060 to 0.8289 MRR@6, indicating that the student benefits from being trained on rankings sampled from its own policy rather than only matching a fixed teacher supervision distribution.

This pattern is consistent with the central interpretation of the paper. Offline KD transfers the teacher's preferences efficiently, but only on the teacher-supported ranking distribution. In contrast, on-policy distillation exposes the student to its *own* ranking errors and asks the teacher to evaluate those rankings. The result is not a large in-distribution gain over offline KD, but a more robust student under distribution shift. For this reason, we treat A3 not as a universally better optimizer than A1, but as a more generalization-oriented distillation strategy.

F Expanded Validation Visualizations Across All Baselines

This section provides subset-level comparisons across all major baselines: Qwen3-4B, Base-1B, BGE v2 M3, Teacher GRPO, Cohere v3.5, Cohere v4-fast, Jina v2, and A3 Distilled-1B. These plots complement the aggregate tables by showing where models differ across datasets.

Figure 10 summarizes per-subset performance across all models. Figure 11 reports win/tie/loss counts against A3 Distilled-1B across the 8 validation subsets. Figure 12 shows the overall shape of each model's subset-level profile, and Figure 13 shows correlations among those profiles.

F.1 Per-Dataset Human–Judge Agreement

To better understand whether judge–human agreement is concentrated in a small subset of the evaluation data, we report agreement broken down by dataset in Figure 14. Agreement remains high across most datasets, with especially strong alignment on InstructIR, InfIR LeetCode, InfoSearch, and FollowIR. The lowest agreement appears on InfIR Robust04, but this bucket is also small and does not change the overall conclusion that the judge is broadly aligned with human relevance judgments.

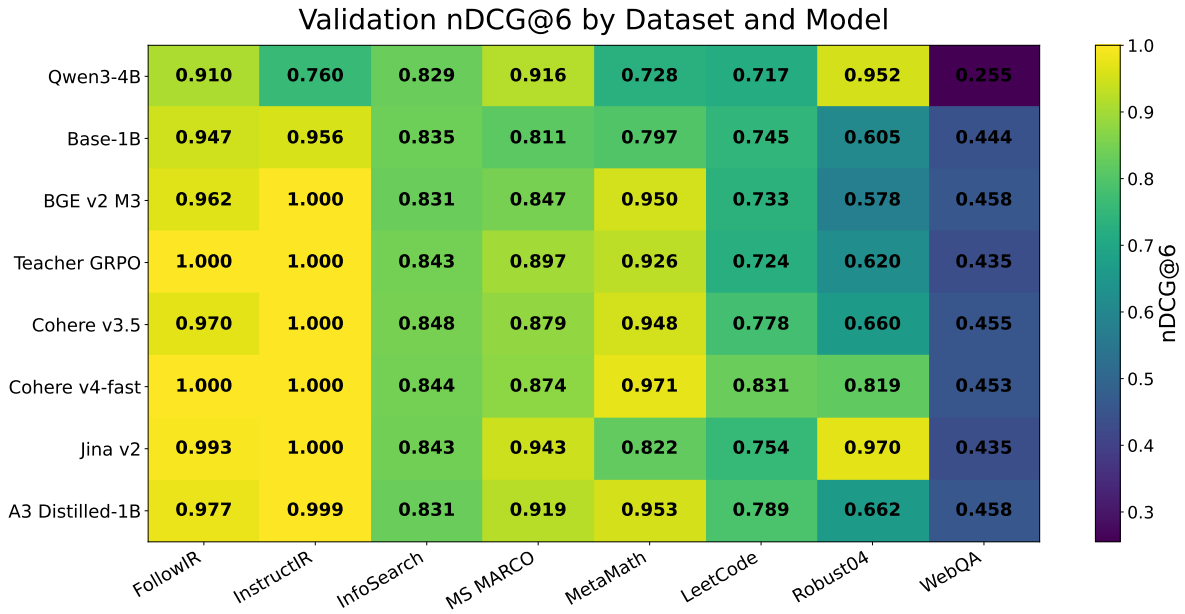


Figure 10: Validation nDCG@6 by dataset and model across all major baselines. The full heatmap makes clear that the distilled student is competitive across a diverse range of subset types, while different baselines exhibit distinct strengths on specific domains.

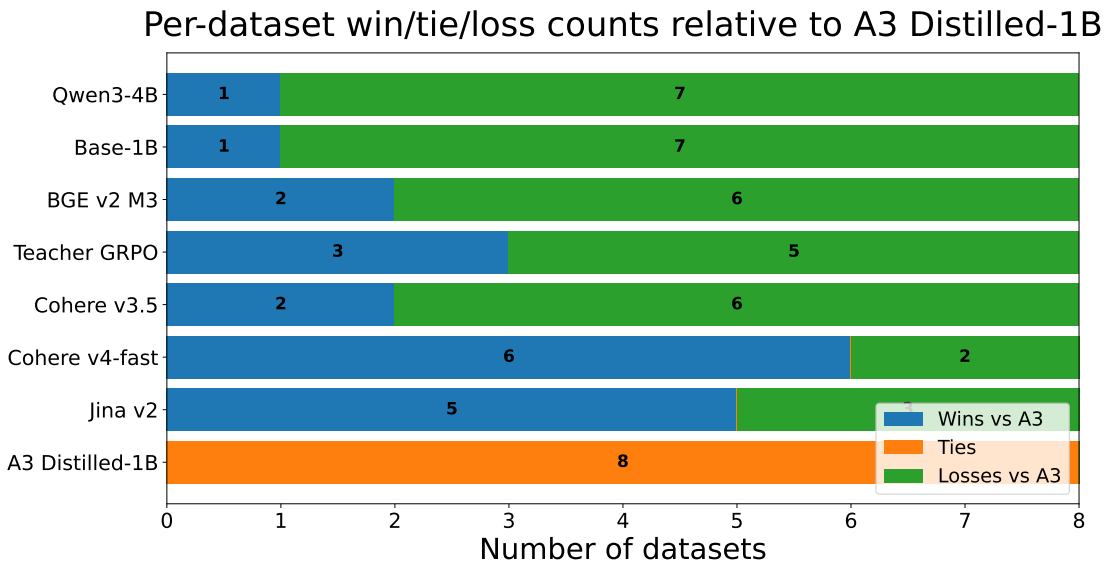


Figure 11: Per-dataset win/tie/loss counts relative to A3 Distilled-1B on the validation benchmark. The distilled student has one of the strongest overall profiles, although different baselines still win on specific subsets.

Comparison / Dataset	Query + Instruction	A3 top-3	Other model top-3	Takeaway
A1 vs A3 *InfIR_robust04*	**Query:** Explore the history and evolution of the property market and its key figures. **Instruction:** Look for articles or blog posts discussing the rise of property developers, their influence, and the changes in the property market over time.	**1.** "history of the property market. Developers became rich, influential and famous . . . " (**rel=1**) 2. "The rapid rise of property development in urban areas has led to a complex . . . " (rel=0) 3. "The history and evolution of the property market is a complex narrative . . . " (rel=0)	**1.** "The rapid rise of property development in urban areas has led to a complex . . . " (rel=0) 2. "The history and evolution of the property market is a complex narrative . . . " (rel=0) 3. "history of the property market. Developers became rich, influential and famous . . . " (**rel=1**)	A3 promotes the only clearly instruction-aligned passage to rank 1, whereas offline KD under-ranks it behind broader but less targeted property-market summaries.
Teacher vs A3 *InfoSearch*	**Query:** Wildlife flourishing in uninhabited areas around Fukushima? I would value a blog entry discussing this matter. **Instruction:** Determine whether a document is relevant to the query.	**1.** "Wildlife Thriving in Fukushima's Uninhabited Zones: A Decade After the Disaster . . . " (**rel=1**) 2. "Nearly a decade after Japan's Fukushima nuclear disaster, researchers have discovered . . . " (rel=0) 3. "User123: I was really surprised to hear that wildlife is flourishing . . . " (rel=0)	**1.** "User123: I was really surprised to hear that wildlife is flourishing . . . " (rel=0) 2. "Nearly a decade after Japan's Fukushima nuclear disaster, researchers have discovered . . . " (rel=0) 3. "Wildlife Thriving in Fukushima's Uninhabited Zones: A Decade After the Disaster . . . " (**rel=1**)	A3 correctly prioritizes the blog-style document that matches both topic and form, while the teacher prefers weaker near-miss passages and buries the relevant one at rank 3.
A6 vs A3 *MAIR_ArguAna*	**Query:** More women in the labour market leads to higher GDP . . . **Instruction:** Given a claim, find documents that refute the claim.	**1.** "gender house believes gender quotas EU are advantageous economies member states . . . " (**rel=1**) 2. "Most corporations, in almost every country on the earth will not even offer . . . " (rel=0) 3. "Slum dwellers in Nairobi are shown to pay high rents . . . " (rel=0)	**1.** "Most corporations, in almost every country on the earth will not even offer . . . " (rel=0) 2. "Slum dwellers in Nairobi are shown to pay high rents . . . " (rel=0) 3. "gender house believes gender quotas EU are advantageous economies member states . . . " (**rel=1**)	Under OOD argumentative retrieval, A3 moves the true counter-argument to rank 1, while the hard-label variant overweights semantically related but non-refuting distractors.
A2 vs A3 *InfIR_msmarco*	**Query:** dame mas gasolina meaning **Instruction:** Relevant documents should explain the meaning of the phrase, including its literal translation and contextual interpretation.	**1.** "Dame Mas Gasolina. It actually means gasoline but in this song it means that . . . " (**rel=1**) 2. "The phrase 'dame mas gasolina' literally translates to 'give me more gasoline' . . . " (rel=0) 3. "The phrase 'dame más gasolina,' while literally meaning 'give me more gasoline' . . . " (rel=0)	**1.** "The phrase 'dame más gasolina,' while literally meaning 'give me more gasoline' . . . " (rel=0) 2. "The phrase 'dame mas gasolina' literally translates to 'give me more gasoline' . . . " (rel=0) 3. "Dame Mas Gasolina. It actually means gasoline but in this song it means that . . . " (**rel=1**)	A3 places the best contextual explanation at rank 1, whereas the off-policy variant prefers generic translation-like distractors and under-ranks the most complete answer.

Table 13: Representative qualitative comparisons drawn from the automatically mined example pool. Each case is a rank-1 flip where A3 promotes the relevant document above the competing model. Together, the examples illustrate four distinct effects discussed in the paper: improvement over offline KD, correction of teacher ranking errors, greater robustness than hard-label training, and better ranking than the off-policy student variant.

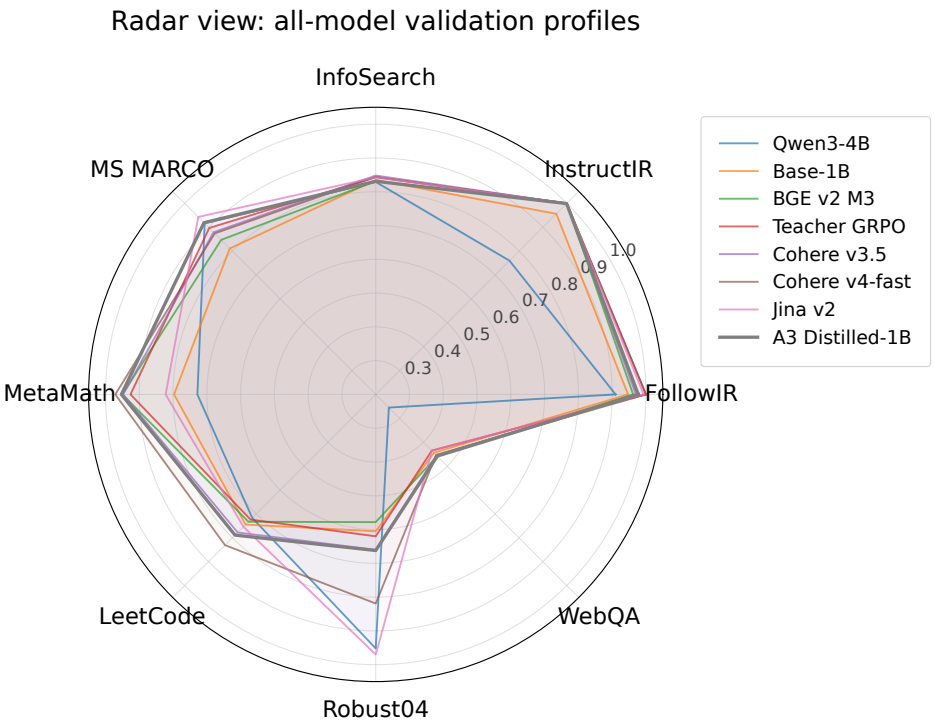


Figure 12: Radar view of validation performance profiles across all major baselines. This visualization emphasizes the overall shape of each model's strengths and weaknesses across the 8 validation subsets.

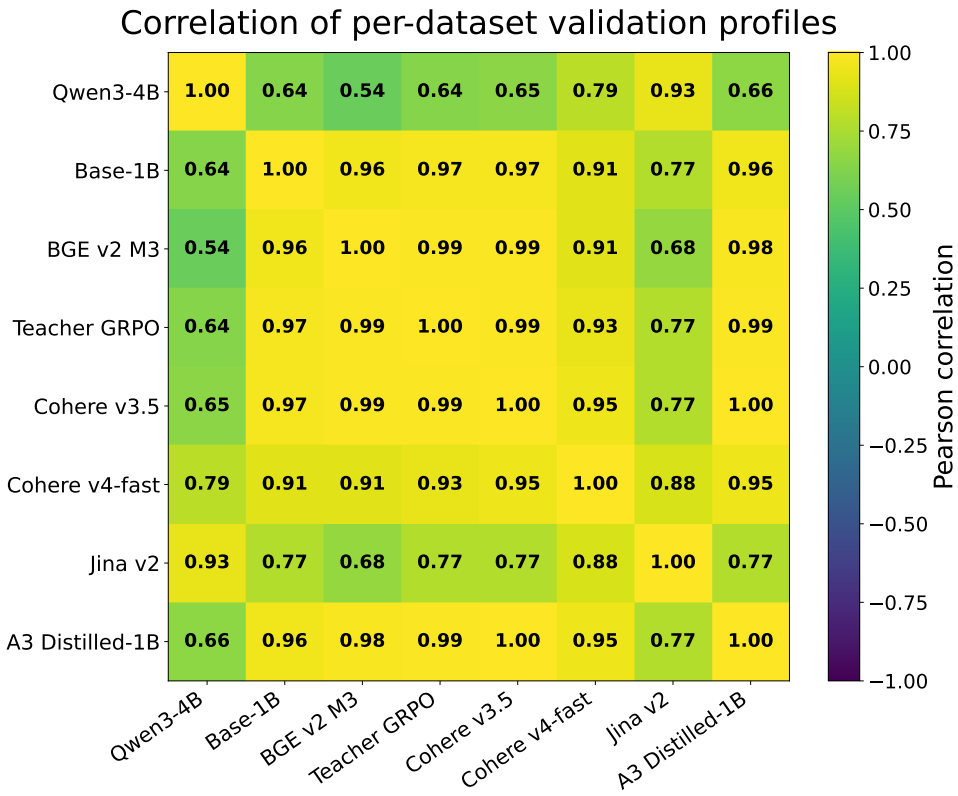


Figure 13: Correlation of per-dataset validation profiles across all major baselines. Higher correlations indicate models that behave similarly across subsets, even when their aggregate scores differ.

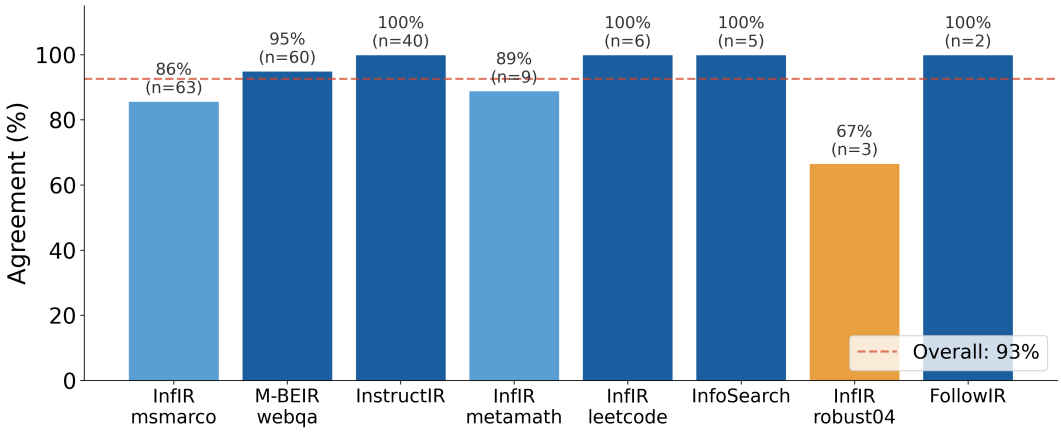


Figure 14: Per-dataset agreement between the LLM judge and human annotations. The dashed line shows the overall agreement rate of 92.6%. Agreement is high across most datasets, with some variation in smaller buckets.