EQUAL RANKING QUALITY, DIFFERENT DECISIONS: MEASURING AND REDUCING ORDER DEPENDENCE IN LLM SCORERS

Anonymous authors
Paper under double-blind review

ABSTRACT

In passage reranking, response ranking and multi-document question answering, LLMs can score several candidate documents or responses together in one prompt, each still receiving its own score. Such scorers are selected on ranking quality, but their scores determine a decision: what a score threshold retains, a reader answers, or which chosen/rejected pair enters preference training. Because the candidates share that prompt, reordering them changes their scores. The same query over the same candidates should still yield the same decision. However, equal ranking quality does not imply equal decisions: on passage reranking, five trained scorers within 0.010 nDCG@10 retain sets that overlap by only 0.66–0.84 when reordered. No prompt-time change we test resolves that dependence: the only one that improves ranking quality does not measurably improve decision stability. We introduce order-consistency SFT (OC-SFT), which attenuates it in the weights by penalizing disagreement between a candidate's scores across orderings. It holds ranking quality and leads every decision-stability measure among trained scorers on all three tasks. It is also more stable on 12 base models than order-averaged distillation, which trains on labels averaged across permutations. One OC-SFT permutation retains sets that overlap more than ten averaged off-the-shelf permutations. A comparison of such scorers should therefore report what a threshold retains and a reader answers, not ranking quality alone. Code is available at `anonymous.4open.science`.

1 INTRODUCTION

A language model that scores candidate documents or responses is rarely the last component of the system it serves. A threshold admits documents above a cutoff into a review queue or generation context, a reader answers from the highest-scoring few, and a preference model's scores determine a chosen/rejected response pair used to train another model (Cui et al., 2024). These three consumers act on the scores, so the output is a decision rather than a ranking, and reordering the same candidates for the same query should not change it. Ranking quality is the property such scorers are selected on, and two that match on it can still differ in every one of those three decisions.

Such a scorer must choose how many candidates share a prompt, and the standard designs each forgo either the per-candidate score or the shared context. Pointwise cross-encoders return a per-candidate value that a threshold can consume, but score one candidate per forward pass (monoT5, Nogueira et al., 2020; RankLLaMA, Ma et al., 2024). Generative listwise rerankers condition on several candidates jointly, but emit a ranked list and no per-candidate score (RankGPT, Sun et al., 2023; RankZephyr, Pradeep et al., 2023a). Batched pointwise scoring, which we study, retains both (Korikov et al., 2025), and its shared prompt also lowers per-query latency, since the instruction, query and grading rubric are encoded once per call rather than once per candidate.

Because the candidates occupy a shared prompt, the score assigned to one is conditioned on the entire prompt rather than on that candidate alone. We term this *presentation dependence*. Language models (LMs) also attend unevenly across long inputs (Liu et al., 2024), which induces sensitivity to candidate order wherever candidates are scored jointly (Zeng et al., 2025; Hou et al., 2024). Off the shelf, the order decides whether a preference model's top choice is right on 65.4% of queries across ten random permutations, so it also decides which response becomes a training label (App. D.6). However, order is not the only such property: with order held fixed, a candidate's score still varies with which other candidates share its prompt and with how the prompt is worded (§3).

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

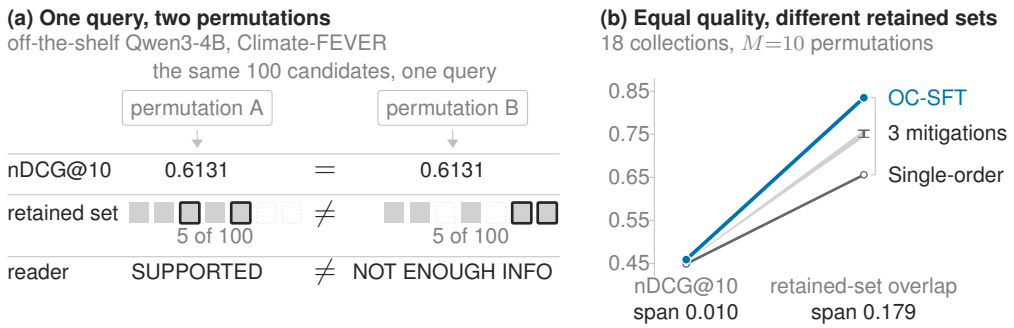


Figure 1: **The same candidates reordered give the same nDCG@10 but a different decision.** (a) Each chip is one document, in the same position in both rows rather than by rank: solid where retained, dotted where not, outlined where they disagree. (b) Five trained scorers; values in Table 1.

None of the prompt-construction changes we test removes this dependence on order, and the established remedy is to score the same pool under several permutations and average (Korikov et al., 2025), which multiplies the serving cost by the number of permutations. Nor does fixing the order remove the underlying dependence: it selects one presentation-dependent output, and first-stage order is not consistently best (App. C.1). Order stability is already reported beside ranking quality for rerankers (Bito et al., 2026) and beside accuracy for LLM judges (Norman et al., 2026), but its effect on the decisions is not. Reordering leaves both the query and the candidate set unchanged, yet each of the three consumers above can produce a different output, and we measure how often they do. Figure 1 shows one query where they do, and that ranking quality does not register it: nDCG@10 reads only the order of the top ten, while a threshold reads the absolute scores.

The question is whether training can reduce what prompt construction cannot. Because a threshold consumes a per-candidate score, we fix how scores are read and vary only how the scorer is trained. Of the objectives we compare (§4), the one whose decisions change least is our order-consistency supervised fine-tuning (OC-SFT), which penalizes the scorer's own disagreement across permutations.

We evaluate on three tasks that score a candidate set in one prompt. Passage reranking is the primary setting across 18 collections and 12 bases in three families. Multi-document question answering (QA) and response ranking test whether the dependence is a property of the shared prompt or of reranking alone (§5). What the shared context adds in quality differs by task and by collection: it helps on QA, not on passage relevance, and on some response-ranking collections (§6.4).
We make the following contributions:

1. We measure what the consumers decide rather than how the ranking moves, and find that equal ranking quality does not imply equal decisions. Five trained scorers within 0.010 nDCG@10 span 0.66–0.84 in retained-set overlap under reshuffling. A comparison should therefore report decision stability, which a stability metric's permutations are enough to measure (§6.1).
2. Fixes applied before training do not improve the decisions. Round-robin partitioning gains 0.052 nDCG@10, while logit calibration improves neither quality nor retained-set overlap measurably. Neither fix reaches any of the three consumers, so we train the score not to depend on order (§6.2).
3. Training improves decision stability without inference-time averaging. At similar ranking quality, OC-SFT is the most stable trained objective on all three tasks; its reader-answer flip rate is 0.125, versus 0.149–0.164 for three other order-aware objectives (§6.3).

2 Related Work

A candidate's score in a shared prompt depends on its position and on which other candidates accompany it. Long-context models attend unevenly across their input, which produces the U-shaped "lost in the middle" effect (Liu et al., 2024). Zeng et al. (2025) measured position sensitivity across retrieval collections and introduced the index we adapt to the rank level (§3). The composition of the prompt exerts a comparable effect, since a candidate's estimated relevance shifts when the accompanying candidates change (Huang et al., 2026). Both channels are documented in reward models and LLM judges that score responses side by side (Malik et al., 2026; Frick et al., 2025; Liu

et al., 2025). Swapping the order alone can flip a two-candidate verdict (Wang et al., 2024), and the flip rate tracks the quality gap between the two (Shi et al., 2025).

Two further channels originate in the construction of the prompt rather than in the candidate set. Partitioning a rank-sorted pool across several prompts renders scores from different prompts not directly comparable, which partitioned listwise reranking addresses by fusing or recalibrating across windows (Parry et al., 2024; Ren et al., 2025). Where scores are extracted from a fixed answer skeleton, its wording is a further design choice, and no position-bias measurement we know of controls for it (§6.2). Attempts to remove this dependence differ in which stage bears the cost.

Inference-time fixes require no training: batched self-consistency averages per-document scores over permutations (Korikov et al., 2025), extending permutation self-consistency (Tang et al., 2024; Wang et al., 2023) and its pairwise form in judging (Wang et al., 2024). TourRank ensembles tournament groupings (Chen et al., 2025), and Huang et al. (2026) sample contexts adaptively. Stable-RAG decodes from the center of several retrieval orders downstream of the ranker (Zhang et al., 2026b), while CapCal (Lv et al., 2026) and CalibraEval (Li et al., 2025) calibrate instead. Resampling repeats its cost per query; calibration pays once, in the data that fits the correction. **Trained fixes** incur it once: DebiasFirst (Qiao et al., 2026) combines position-aware augmentation with inverse-propensity weighting, RankVicuna (Pradeep et al., 2023b) distills from shuffled orders, Brown & McIlroy (2025) fine-tune on the order-stripped inputs of McIlroy-Young et al. (2024), and Zheng et al. (2026) reward consistent decisions with reinforcement learning. The cost is supervision and one training run, and the outcome is attenuation, not a guarantee (Bito et al., 2026). Penalizing disagreement between views of the same input is a standard consistency-regularization device of semi-supervised learning (Tarvainen & Valpola, 2017; Liang et al., 2021), and Xiang et al. (2024) and Zhang et al. (2026a) apply it across permuted demonstrations and across prompt phrasings. Both stabilize a discrete answer choice, which no threshold consumes, whereas we stabilize a continuous per-document score at the cutoff a threshold applies to it, and we measure what that threshold then retains (§6.1). **Architectural fixes** remove the dependence by construction: SetRank (Pang et al., 2020) encodes the candidates as a set rather than prompting an LM, keeping the shared context, while LLM-side variants restrict cross-candidate attention to obtain order invariance in a single pass (Wang et al., 2025b; McIlroy-Young et al., 2024; Bito et al., 2026). The cost falls on the mechanism itself, since that restriction reduces the cross-candidate information the shared prompt was adopted to supply.

Outside these families, rerankers are trained for quality rather than for presentation dependence, on large hard-negative-mined corpora (Pradeep et al., 2023a; Wang et al., 2025a; Li et al., 2026), and reward models for response selection on large preference corpora (Liu et al., 2026; Jiang et al., 2023).

We attenuate the dependence in training, which leaves the serving pass at one permutation and the shared prompt unrestricted. Stability is reported beside accuracy for judges (Zheng et al., 2023; Shi et al., 2025; Norman et al., 2026) and beside ranking quality for reranking (Bito et al., 2026), in each case on the scorer's own output. We measure it instead on what the consumers of that output produce, and as a change across permutations rather than a value under a single permutation (§6.1).

3 Presentation dependence and how we measure it

A scorer that reads several candidates in one prompt assigns each of them a score that depends on the rest of the set. We score a pool of candidates in windows of B per prompt, partitioning the pool where it exceeds B, and sort it by the resulting scores. In reranking, a first stage's top 100 becomes five contiguous windows of $B = 20$, the window size of listwise rerankers (RankGPT, Sun et al., 2023; RankZephyr, Pradeep et al., 2023a). A candidate's score depends on its window assignment, on which others share that window, and on its position among them (App. C.1). Reshuffling before partitioning moves all three, and with them the ranking[1]

Presentation channels. Four properties of the prompt can vary while the query and the candidate pool are held fixed: (i) the order in which candidates are presented, (ii) the rule partitioning that order into windows, (iii) the markers that label the slots, and (iv) the wording of the answer skeleton from which scores are extracted. We call one setting of all four a *presentation*, and a scorer's sensitivity to it *presentation dependence*; pool composition is separate, since it changes the candidate set and lies

[1]This is not specific to our readout setup. RankZephyr generates a ranked list and jina-reranker-v3 scores 100 candidates in a single context, and both are order-sensitive at their native geometries (App. H.3).

outside the prompt's control. Except where stated, we vary the order and hold the rest fixed. The partition rule (ii) is set when the prompt is built, so a different rule removes its cost. Where the pool fits one window, as for QA, nothing is partitioned. The markers (iii) and the skeleton (iv) are set at the same point and held fixed by default, since no skeleton setting is neutral; we vary each separately (§6.2; App. C.1). None of these removes the order channel (i), where the decisions vary.

Measuring order dependence. We measure the order channel with τ-PSI, adapting the index of Zeng et al. (2025) to the rank level: for each instance we score its candidate set under $M = 10$ random permutations and take the induced rankings $r_1, \ldots, r_M$. We then average their pairwise Kendall correlation $\tau(r_i, r_j)$:

$$\tau\text{-PSI} = \frac{1 - \text{mean}_{i<j}\tau(r_i, r_j)}{2},$$

which is 0 when every order produces the same ranking, 0.5 when rankings are uncorrelated, and 1 when reversed. A permutation-invariant scorer treats its input as a set and attains 0, so τ-PSI is a distance from set-like behavior and lower is more stable. We randomly shuffle candidate order to model ordinary deployment variation, not a worst case; index-driven pool changes are evaluated separately (§6.3). τ-PSI is conditional on the partition rule, markers and skeleton. We hold all three constant, so a difference in τ-PSI is a difference between scorers.

Measuring what the consumers produce. For each consumer, we report output quality and how often the output changes, since a scorer could hold it fixed by making it worse. Throughout, change is the rate at which two permutations disagree, averaged over pairs, and independent of how many are drawn. τ-PSI is not a substitute for those measures: it measures movement in the ranking rather than in what a consumer produces, so the two need not rank scorers the same way (App. F).

Two separate quantities. We write the scorer output for a query and candidate set $x = (q, D)$, at width B and under a candidate order π, aligned back to candidate identity, as

$$f_B(x, \pi) = \underbrace{\mu_B(x)}_{\text{order-marginal}} + \underbrace{\delta_B(x, \pi)}_{\text{order residual}},$$

with $\mu_B(x) = \mathbb{E}_\pi[f_B(x, \pi)]$ and $\mathbb{E}_\pi[\delta_B] = 0$. The *order-marginal* is the score a permutation-invariant scorer at that width would emit. The *residual* is what remains, and it varies when only the input order changes; τ-PSI measures the residual through its effect on the ranking. Batched self-consistency (BSC) estimates the order-marginal by averaging a scorer's output over permutations (Korikov et al., 2025). The objective of §4 targets the same order-invariance at one permutation, by penalizing the residual's empirical variance while leaving μ_B unconstrained. Training should therefore reduce order instability before any serving-time averaging (§6.3).

Width and scope. We call $\mu_B - \mu_1$ the *width effect*, the difference between scoring a candidate at width B and scoring it in isolation. A pointwise scorer sets $B = 1$, where the residual and the width effect are both zero by construction, so it avoids the order channel and forgoes an interaction whose sign varies by collection (App. E). The decomposition runs over permutations of a fixed candidate set, so replacing a candidate changes x rather than π and falls outside it. Whether an objective targeting δ_B reaches that perturbation is measured in §6.3.

4 Where to address order dependence

§3 locates the order dependence in the residual δ_B. Holding the readout and teacher fixed, we ask whether penalizing that residual changes what the consumers decide at a single serving permutation.

Score readout. A generic instruction-tuned chat model does not emit the per-candidate score that batched pointwise scoring requires, so we extract the score without generating any text. The prompt supplies a grading instruction, the query, and the B candidates that share the prompt, each tagged `[i]`, and requests one integer grade in $\{0, 1, 2, 3\}$ per candidate. Rather than sampling those grades, we append a fixed answer skeleton with a placeholder grade token per candidate, so every grade position is known in advance and no decoding temperature enters. One forward pass over the prompt and skeleton yields the model's distribution over the four grade tokens at each position, renormalized over those four. We take the expected grade $E[g] = \sum_g g\, P(g)$, normalized to $[0, 1]$, as the score, chosen by comparing readouts on untrained bases. One prompt thus yields all B scores. Unlike a generated ranking, the readout cannot omit or repeat a candidate. The score is continuous and comparable across

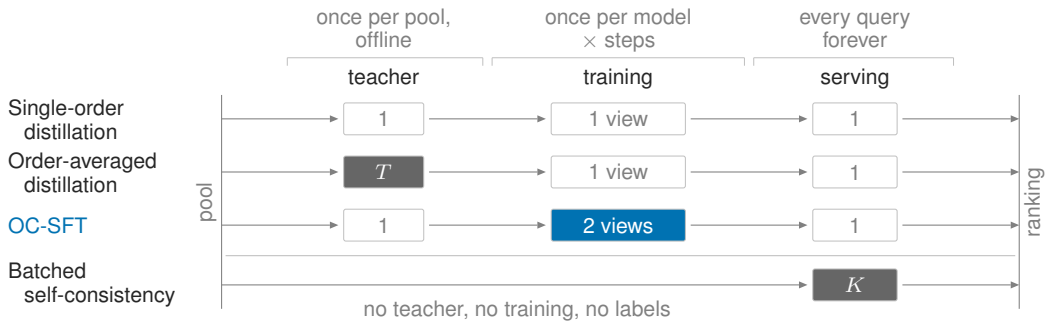


Figure 2: **Order dependence can be addressed at serving, in the labels, or once in training.** Each row is one approach, and the filled marker is where it encounters several permutations; single-order distillation, the main baseline, never sees more than one. T counts permutations of one training instance and K permutations of one query; the two are independent, and OC-SFT's two views are the $N{=}2$ of its objective. We use $T{=}10$ and $K{=}10$; Figure 4 measures how far averaging gets.

instances, so it admits a decision threshold, and the readout is unchanged at any width down to $B{=}1$. We apply the same readout at $B{=}10$ for QA and $B{=}4$ for response ranking training (App. B.1).

Supervision. A teacher scores the candidates of each training instance and produces one target per candidate. Because the readout is the expected grade, the targets are continuous rather than binary relevance labels. We hold the teacher and the training pool fixed, so the variants below differ only in their objectives. By default the teacher is the student's own base model (self-distillation).

Training objectives. The approaches to μ_B differ in where they encounter several permutations (Figure 2): self-consistency at serving time, order-averaged distillation offline in the labels, and OC-SFT in its N views per step. All trained variants distill from these targets, minimizing a squared error between the student's expected grade $s(d)$ and the teacher target $\tilde{y}(d)$, averaged over the candidates of a $B = 20$ window. Single-order distillation, or plain fine-tuning, uses the target from one teacher order and does not target μ_B. Order-averaged distillation uses the batched self-consistency target, the per-candidate average over $T = 10$ shuffled teacher orders precomputed into the label. The label absorbs the averaging, so order-averaged distillation targets μ_B through the teacher.

We introduce **order-consistency SFT** (OC-SFT), which instead constrains the student's own order residual δ_B (§3), following the consistency-regularization principle of semi-supervised learning (Tarvainen & Valpola, 2017; Liang et al., 2021). A permutation-invariant scorer has no such residual, so penalizing it targets invariance in the student rather than through the label. OC-SFT retains the single-order target and scores the same window under N permutations $\pi_1, \ldots, \pi_N$, with per-candidate scores $s_i(d)$, each f_B at π_i restricted to d, and view-mean $\bar{s}(d) = \frac{1}{N}\sum_{i=1}^{N} s_i(d)$. [2] A penalty with weight λ then pulls each view toward that mean:

$$\mathcal{L}_{\text{OC-SFT}} = \underbrace{\text{mean}_d \left(s_1(d) - \tilde{y}(d)\right)^2}_{\text{relevance anchor}} + \lambda \underbrace{\frac{1}{N}\sum_{i=1}^{N} \text{mean}_d \left(s_i(d) - \bar{s}(d)\right)^2}_{\text{order variance: } \delta_B}.$$

Only the first view, itself one of the shuffles, carries the relevance anchor, so more views add no teacher supervision. The second term is the empirical order-variance of the student's own scores, proportional in expectation to $\mathbb{E}_\pi \|\delta_B(x, \pi)\|^2$, so it directly penalizes the within-window residual. The penalty acts on scores and bounds their variance rather than removing it, whereas τ-PSI reads rankings; near a tie, the surviving variance can still move a rank. The objective therefore yields attenuation rather than invariance, in the ranking and in the decisions read off it (§6.3).[3] We use $N = 2$.[4] The tenfold saving over order-averaged distillation is in offline teacher labels and not in training compute.

[2] These shuffles permute slots within a window, so a candidate keeps its companions, whereas a τ-PSI shuffle can change them; the metric therefore perturbs at least as much as the penalty targets.

[3] In group-relative terms, μ_B is the group mean and δ_B the deviation, as in GRPO (Shao et al., 2024); the group varies the arrangement, not the sampled action, so the residual is differentiable and needs no rollouts.

[4] One pair estimates that variance up to a constant absorbed into λ, and we detect no quality change to $N = 8$ (App. D.3). Two views roughly double the training step cost; all trained variants serve one permutation.

Both terms are required: without the relevance anchor the model collapses to a consistent but relevance-free scorer, and at $\lambda = 0$ the objective reduces to single-order distillation. The readout is unchanged across widths, so the weights can serve a different B without retraining (App. G).

5 Experimental setup

Data. We train and evaluate on three tasks that score a candidate set in one prompt. Each has its own training data and teacher labels; the readout and the objective are the same throughout.

Passage reranking is the primary setting. We train on MS MARCO (Nguyen et al., 2016) at $B = 20$, deriving labels from a teacher that grades the top-100 candidates of ∼30K queries under the readout of §4. Evaluation covers 18 collections: five in-domain TREC Deep Learning collections built over MS MARCO (DL19 through DL23, Craswell et al., 2020), 11 out-domain BEIR collections (Thakur et al., 2021), and two internal legal collections from a commercial legal search provider, with the BEIR and legal groups both evaluated zero-shot. The 16 public collections match the suite that recent listwise and distillation rerankers report on. The only two that carry claim-verification labels, SciFact with 300 claims and Climate-FEVER with 1,381, also serve as the downstream collections for verdict flip. **Multi-document QA** scores answer support rather than topical relevance: the HotpotQA distractor setting, ten passages per question with exactly two gold and binary relevance (Yang et al., 2018), trained on 30K questions and transferred zero-shot to 2WikiMultiHopQA (Ho et al., 2020) and MuSiQue (Trivedi et al., 2022). **Response ranking** scores candidate responses for preference or correctness, trained on ∼30.7K UltraFeedback prompts (Cui et al., 2024) with four candidates each. Its five collections each cover a distinct regime, differing in the quality gap between candidates and in whether one can be judged in isolation: RewardBench-2 (Malik et al., 2026), Nectar (Zhu et al., 2023), PPE MMLU-Pro and PPE-MATH (Frick et al., 2025), and RM-Bench (Liu et al., 2025). App. B.2 lists every evaluation collection; scope and limitations are discussed in App. A.

Models and teachers. Qwen3-4B (Yang et al., 2025) is our primary base model, one of 11 dense bases from 1.7B to 32B across three families: Qwen3, Gemma 4 (Gemma Team, 2026) and Granite 4.1 (IBM Granite Team, 2026). Two bases serve as controls, the sparse Mixture-of-Experts Gemma-4 26B-A4B for architecture and Qwen3-Reranker-4B for reranking-specialized initialization. The default teacher is the student's own base model (self-distillation), so the method needs no external model and the reported gains cannot come from a stronger teacher; we also evaluate larger open teachers within and across families (Qwen3, Gemma 4, Granite 4.1) and GPT-5.4. We train with LoRA (Hu et al., 2022) throughout, selecting checkpoints and hyperparameters per model and variant on a held-out split (App. B.3); base and teacher models are varied in App. H.4.

Baselines. All reranking baselines use the same 100 BM25 candidates as our trained models; we repeat the comparison on a dense (BGE-base-en-v1.5, Xiao et al., 2024) and a learned-sparse (SPLADE++ ED, Formal et al., 2022) first stage (App. D.7). The inference-time comparison is batched self-consistency (Korikov et al., 2025), the K-permutation ensemble, together with CapCal (Lv et al., 2026), a content-agnostic logit calibration applied to the off-the-shelf base. As a trained mitigation we adapt DebiasFirst (Qiao et al., 2026), position-aware augmentation with inverse-propensity weighting, to batched score regression. Alongside it we report permutation augmentation, as in RankVicuna (Pradeep et al., 2023b), which retains the N shuffled views and anchors each to the teacher target.

We also report the purpose-built jina-reranker-v3 (Wang et al., 2025a) at our $B = 20$ geometry and a frontier prompted scorer, GPT-5.4, on reranking, QA, and response ranking; nearly half its candidate pairs tie, so we omit its τ-PSI. App. H.3 additionally evaluates RankZephyr (Pradeep et al., 2023a) and mxbai-rerank-large-v2 (Li et al., 2026) at their native geometries. On response ranking we compare against the pointwise reward model Skywork-Reward-V2-Qwen3-4B (Liu et al., 2026) and the pairwise PairRM (Jiang et al., 2023); neither sees a candidate order, so we report quality only (App. E.2).

Ranking metrics. We report nDCG@10 for quality and τ-PSI for order instability (§3); truncation at 10 affects only reranking's 100 candidates, since QA ranks its ten candidates in full. Response ranking instead reports nDCG@1, the normalized relevance gain of the top-ranked response.

Decision metrics. For the threshold results, we freeze a per-collection F1-tuned cutoff, fitted on half of a collection's queries and measured on the other half. Reshuffling therefore changes what the filter admits and not how it was tuned. For that retained set we report its F1 and its mean pairwise Jaccard, the overlap between the sets retained under different permutations. Downstream, two frozen

Table 1: **Quality saturates on all three tasks; what the consumers decide does not.** Each task block gives quality, order instability, and its consumer's decision: retained-set overlap (Jacc.), answer flip and chosen/rejected-pair flip. **Bold** marks a Holm-adjusted lead over the other trained variants on all three seeds (App. B.4); no quality column qualifies (§6.1). Basis: Qwen3-4B, $M=10$ random permutations, trained rows self-distilled 3-seed means; BSC is batched self-consistency (Korikov et al., 2025) over the off-the-shelf scorer. Collections and per-task widths are in App. B.

	Passage reranking (18)			Multi-doc QA (3)			Response ranking (5)		
Variant	nDCG@10↑	τ-PSI↓	Jacc.↑	nDCG@10↑	τ-PSI↓	Ans.↓	nDCG@1↑	τ-PSI↓	Pair↓
Untrained									
Off the shelf	0.370	0.298	0.439	0.911	0.224	0.221	0.655	0.345	0.877
CapCal	0.372	0.293	0.427	0.911	0.222	0.217	0.657	0.338	0.874
Round-robin	0.422	0.297	0.443	0.911	0.224	0.221	0.658	0.347	0.878
BSC ($\times$10)[a]	0.465	0.180	0.707	0.946	0.143	0.157	0.716	0.184	0.636
External									
jina-reranker-v3	0.447	0.177	0.667	0.949	0.163	0.172	0.479	0.226	0.685
GPT-5.4	0.468	—[b]	0.707	0.972	—[b]	0.094	0.726	—[b]	0.489
Trained									
Single-order	0.449	0.209	0.656	0.951	0.159	0.177	0.684	0.333	0.869
Order-averaged	0.455	0.130	0.743	0.956	0.124	0.149	0.693	0.228	0.724
DebiasFirst	0.454	0.128	0.759	0.955	0.147	0.164	0.694	0.228	0.718
Perm. aug.	0.455	0.129	0.760	0.955	0.148	0.162	0.696	0.223	0.707
OC-SFT	0.459	**0.083**	**0.835**	0.961	**0.096**	**0.125**	0.701	**0.201**	**0.661**

[a] Its order-instability and decision cells compare four independent ten-permutation ensembles, the unit such a system serves; every other row compares single permutations, at a tenth of the inference.
[b] Nearly half its candidate pairs tie. We use an order-independent tie-break; stable sorting raises answer/pair flip from 0.094/0.489 to 0.200/0.851. τ-PSI credits fixed tie orders to the model, so we omit it.

readers consume the scorer's top 5 under the same $M=10$ permutations, Granite-4.1-8B throughout and Qwen3-4B as a second reader. For them, we report verdict accuracy and *verdict flip* on claim verification, and exact match, F1 and *answer flip* on QA. A preference model's scores determine a chosen/rejected response pair, for which we report *pair flip*, how often that pair changes. Each consumer is held fixed, so any change comes from the scorer alone.

Aggregation. Task means equally weight 18 reranking, 3 QA and 5 response-ranking collections; τ-PSI and flip metrics use $M=10$ random permutations. Trained rows average 3 seeds, otherwise seed 42; intervals use paired 95% bootstraps over collections then queries.

6 Results

6.1 Decisions at matched ranking quality

Ranking quality does not predict how far a scorer's decisions change under reshuffling. The five *trained* variants in Table 1 differ by at most 0.010 nDCG@10 on passage reranking, less than a permutation moves a single query's score under any one of them (Figure 3a). Across the $M=10$ random permutations of each pool, retained sets overlap pairwise by 0.656 under single-order distillation and 0.835 under OC-SFT, a difference more than an order of magnitude larger than that quality spread. Among the seven single-permutation systems in Table 1 within 0.022 nDCG@10, different systems take the best quality and the best retained-set overlap on 11 of the 18 reranking collections (App. F.2). A threshold reads absolute scores; nDCG@10 reads only the discounted order of the top ten. So a small score change near the cutoff admits or drops a document, while similarly graded candidates trading places barely moves nDCG@10. A comparison of such scorers should report what a threshold retains or a reader answers, not the ranking metric alone.

Two published systems lead on a quality measure while reproducing less of their own output than the scorer they outrank, sharing neither our base model nor our supervision. GPT-5.4 leads every variant we trained on nDCG@10 while reproducing less of its retained set than OC-SFT (Table 1).

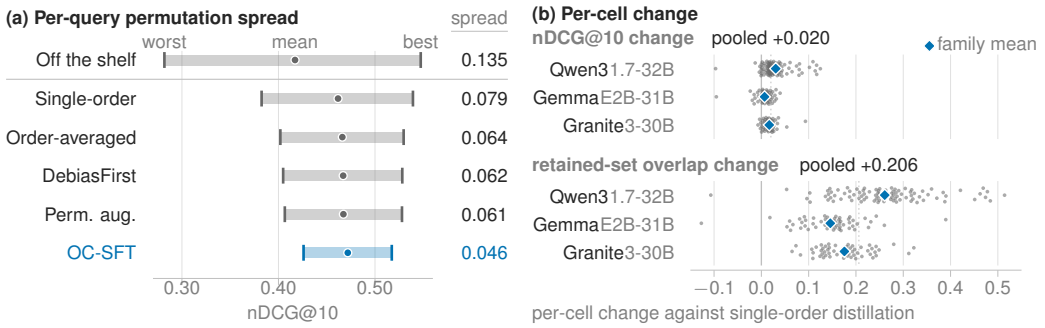


Figure 3: **Ranking quality moves across permutations and barely with the method; quality and retained-set overlap separate in every family.** (a) Qwen3-4B per-query nDCG@10 over $M=10$ random permutations on 18 collections; *spread* is mean minus worst. (b) Dots are 198 base-collection cells over 11 dense bases; overlap at a per-collection F1-tuned cutoff, *pooled* is their mean.

`jina-reranker-v3`, a purpose-built reranker rather than a frontier model, leads OC-SFT on retained-set F1 while trailing it on nDCG@10, and its retained sets overlap less, 0.667 against 0.835.

6.2 Channels fixed at prompt construction

Fixes applied before training do not improve the decisions. The contiguous partition of a rank-sorted top 100 (§3) makes each window internally uniform on a four-grade scale, so equal grades in different windows denote different relevance levels and the merged ranking conflates them. Round-robin assignment distributes consecutive first-stage ranks across distinct windows and recovers 0.052 nDCG@10 on reranking, without additional permutations (Table 1, *untrained* rows). However, it leaves decision stability where it was: off the shelf, neither retained-set overlap, verdict flip on either collection, nor pair flip changes measurably. A post-hoc logit calibration (Lv et al., 2026) improves neither quality nor retained-set overlap measurably, and leaves τ-PSI near the off-the-shelf level. Rescaling across windows (Ren et al., 2025) cannot correct the partition effect at all, since the windows do not differ by an offset. We report under contiguous windows, the conventional partition. Under round-robin the trained gain is smaller and still positive, and the partition rule moves no variant's τ-PSI by more than 0.006. Detailed results are in App. D.

Scores are read from a fixed grade slot, so the placeholder in every unscored slot is part of the presentation. We adopt `Grade: 0`, the in-format placeholder with the best average ranking quality across our three families, and hold it fixed; no other in-format value lowers τ-PSI distinguishably (App. B.1). We are aware of no position-bias measurement, including the index we adapt, that reports its sensitivity to this choice; such numbers are therefore in part a property of the skeleton behind them. None of the alternatives tested removes the order dependence that moves the decisions.

6.3 What training changes for the consumers

The penalty moves what the consumers decide far more than it moves ranking quality. Order instability on passage reranking falls from 0.209 under single-order distillation to 0.083 under OC-SFT, and the retained set, the reader's output and the selected pair all become more reproducible (Table 1). On quality alone the two variants are close; on retained-set stability they are not. As Figure 3a shows, OC-SFT also narrows per-query score movement below every other variant.

The gain holds model by model and under unseen perturbations, and no control removes it. Quality and retained-set overlap separate the same way per model and collection across 11 dense bases in three families (Figure 3b), and OC-SFT stays below order-averaged distillation in τ-PSI on all 12, including the sparse MoE. The same ordering appears under dense and learned-sparse first stages. Under pool replacement, a perturbation the objective was never trained on, instability under that perturbation is 0.031 against single-order distillation's 0.099. Matching retention at the F1-tuned cutoff leaves the gain in retained-set overlap in place. What the penalty removes is movement at the ranking head, not ties or collapsed scores, and the ordering holds under worst-permutation quality and dispersion. The Climate-FEVER verdict changes less often than under single-order distillation,

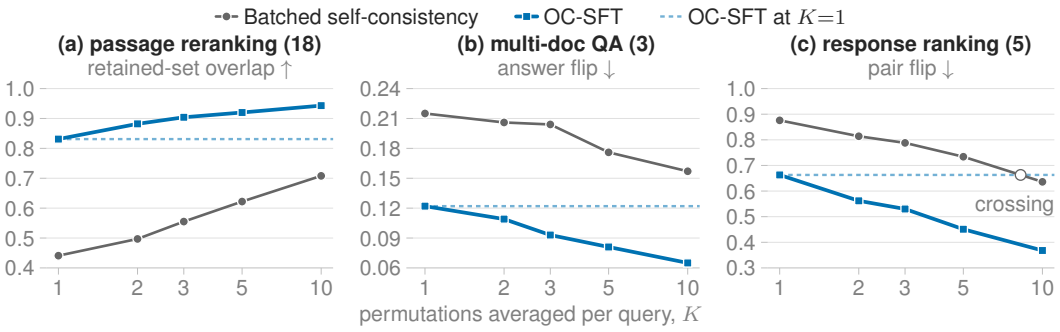


Figure 4: **OC-SFT at K=1 beats BSC at K=10 permutations on two of the three consumers.** Batched self-consistency (BSC, grey; Korikov et al., 2025) reaches OC-SFT's own K=1 (dashed) only on (c), near K=8; averaging keeps improving OC-SFT itself. BSC's K=10 is its Table 1 row.

with no measurable accuracy change. The lower verdict-flip rate persists among queries for which both variants select the same five documents. For detailed results, we refer to Apps. C, D and F.

Permutation augmentation, which trains on the same shuffled views as OC-SFT but omits its explicit consistency penalty, closes about two thirds of the τ-PSI gap, as does DebiasFirst. A generic consistency mechanism closes as much of that gap and still fails the consumer: an EMA mean-teacher closes 71% of it while costing the reader 0.023 exact match (Apps. D and F for full results). The penalty closes the remaining third; shuffled-view training alone does not match OC-SFT's stability.

The designs that do reach the decisions pay for it at every query or every label. A self-consistency ensemble repeats the serving pass on every query (Korikov et al., 2025), and at K=10 permutations it does not reach OC-SFT at K=1 on two of the three consumers (Figure 4). Averaging the trained scorer improves its reproducibility too, so training and averaging compose rather than substitute. Order-averaged distillation pays in the labels instead, needing T=10 teacher passes against OC-SFT's one, and is less stable on 30K labeled queries than OC-SFT on 1K. A stronger teacher buys quality but not stability, so the gain comes from the penalty rather than the supervision (Apps. D and H).

6.4 Serving width and the choice of design

The penalty suppresses order dependence without sacrificing shared context where it helps. Serving one checkpoint at its training width rather than at B=1 gains 0.043 nDCG@10 on multi-document QA. On passage relevance at B=20 the width effect is indistinguishable from zero, and on response ranking the gain is largest on the PPE collections while RM-Bench loses (Apps. E and G).

Which design to serve follows from the serving budget and the task; the shared prompt earns its place in the labels either way. Served one candidate at a time, where both designs are order-free, a pointwise student trained for that width matches the batched checkpoint on passage relevance, 0.466 against 0.459 on an interval covering zero. One batched checkpoint holds quality across 1–30 candidates; the pointwise student degrades at 20. The labels allow no such choice: batched-teacher labels beat pointwise ones at both training widths, by 0.029 nDCG@10 at B_{train}=1 and 0.010 at B_{train}=20, both evaluated at B=1. At B=20, batching holds that quality at 45% lower per-query latency. Detailed comparisons are in Apps. D and G.

7 Conclusion

Scorers a ranking metric cannot separate can still differ in what their consumers decide, so we compared remedies by where their cost falls: at serving, in teacher-label construction, or once in training. Fixes applied before training reached no decision: round-robin partitioning improved ranking quality without making any of the three decisions measurably more reproducible, while logit calibration improved neither quality nor retained-set overlap. The order channel resisted both; OC-SFT, a consistency penalty over shuffled views, attenuated it in the weights from one teacher pass and one serving permutation, while costing no ranking quality and making all three decisions more reproducible. It also lowered order instability on all of the 12 bases we trained. Telling quality-matched scorers apart therefore takes another measurement, and a stability metric's permutations are enough to run it on.

ETHICS STATEMENT

This work studies the reliability of LLM scorers rather than a new capability, since the same candidate set producing different rankings under reordering is a reproducibility concern in high-stakes retrieval such as legal search. We train on MS MARCO, a public web-search dataset, and evaluate on public TREC-DL and BEIR collections with two proprietary legal collections from a commercial legal search provider, reported as Legal-A and Legal-B. Both contain professionally authored content and attorney relevance judgments. We collected no new annotations and release no raw queries, documents, or annotator identifiers from them. They were evaluated under institutional data-governance controls. TREC-NEWS and Robust04 are distributed and reported under NIST terms. One closed model, GPT-5.4, appears as a comparison scorer and teacher, accessed through its provider's API under the terms of service. Its weights and training data are not inspectable and the provider can retire the served version, so the GPT-5.4 comparisons are less reproducible than the rest. The method requires no closed model, because the default teacher is the student's own base model and every base we train is open-weight. We evaluate closed teachers only as a comparison, and an open teacher recovers most of their advantage without GPT-5.4 labels (App. H.4).

REPRODUCIBILITY STATEMENT

App. B.1 gives the scoring readout and reranking prompt template, including the scored answer skeleton. App. D.3 covers silver-label construction and the consistency-weight selection protocol. §5 defines τ-PSI and states the permutation budget and the resampling unit for intervals, with the evaluation collections in App. B.2. App. B.3 lists the shared training configuration and principal task-specific settings: base model, LoRA settings, optimizer and schedule, training width, consistency weight, selection rule, and hardware. Per-collection reranking results are in App. H, with the per-task, fixed-width and perturbation-specific breakdowns in Apps. C.1, E and G.

We release the τ-PSI implementation and the evaluation harness, the scoring prompt and scored skeleton of App. B.1, and the training and evaluation code, including the inference script. We also release the silver labels behind the primary result, the LoRA adapters for the 11 dense-base students of §5, and the full evaluation configurations and their commands. The repository is at `anonymous.4open.science/r/order-dependence-llm-scorers`. The full public evaluation suite runs in 60–90 hours on a single L40S-48GB GPU, and one collection such as DL19 in about an hour. Training the student behind the primary result takes about 90 GPU-hours on one 8-GPU A100-40GB node from the released silver labels. The full study, including unreported runs, totals roughly 100,000 GPU-hours.

Legal-A and Legal-B are proprietary, so the 18-collection reranking means in Table 1 and §6 include two collections others cannot score; App. H reports the sixteen-collection public mean alongside every per-collection cell. We accessed GPT-5.4 through an Azure OpenAI `gpt-5.4` deployment, served by model version `gpt-5.4-2026-03-05`. Evaluation calls ran between June and August 2026, and its teacher labels for App. H.4 between May and June. Temperature and top-p are not configurable on this interface, so provider defaults applied. Rescoring byte-identical prefixes in separate inference batches does not return the same number, so a reader recomputing a score-level figure from the released outputs should not expect an exact match (App. G).

AI USE STATEMENT

Large language models are an object of study and a component of the method in this work. This covers the scorers we train and evaluate, the teachers that generate the silver relevance labels, and the prompted closed-model baselines, all described in §4 and §5. They are studied here rather than used to produce the work, and the silver-label generation of App. D.3 is part of the method rather than a synthetic-data tool use.

We used generative AI tools for implementing software and preparing data, for literature search, brainstorming and interpreting results, for drafting the manuscript and editing it for length and readability, and for formatting references. We verified references against primary sources rather than relying on model output, and we checked numeric claims in the text against the floats they cite. We did not use generative AI tools to propose or refine the hypotheses or to formulate the decomposition of §3. The authors reviewed all AI-assisted work, made all scientific decisions, and take responsibility for the final content, including text, claims and artifacts produced with generative AI.

REFERENCES

Ethan Bito, Yongli Ren, and Estrid He. One pass, any order: Position-invariant listwise reranking for LLM-based recommendation (InvariRank). In *SIGIR*, pp. 3625–3629, 2026. doi: 10.1145/3805712.3809952. URL https://arxiv.org/abs/2604.27599. arXiv:2604.27599.

Katrina Brown and Reid McIlroy. Order independence with finetuning. In *ICLR Bi-Align Workshop*, 2025. URL https://arxiv.org/abs/2503.23483. arXiv:2503.23483.

Yiqun Chen, Qi Liu, Yi Zhang, Weiwei Sun, Xinyu Ma, Wei Yang, Daiting Shi, Jiaxin Mao, and Dawei Yin. Tourrank: Utilizing large language models for documents ranking with a tournament-inspired strategy. In *Proceedings of the ACM Web Conference 2025 (WWW '25)*, 2025. doi: 10.1145/3696410.3714863. URL https://arxiv.org/abs/2406.11678. arXiv:2406.11678.

Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Ellen M. Voorhees. Overview of the TREC 2019 deep learning track. *arXiv preprint*, 2020. URL https://arxiv.org/abs/2003.07820. arXiv:2003.07820.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, Zhiyuan Liu, and Maosong Sun. UltraFeedback: Boosting language models with scaled AI feedback. In *ICML*, pp. 9722–9744, 2024. URL https://proceedings.mlr.press/v235/cui24f.html.

Thibault Formal, Carlos Lassance, Benjamin Piwowarski, and Stéphane Clinchant. From distillation to hard negative sampling: Making sparse neural IR models more effective. In *SIGIR*, pp. 2353–2359, 2022. doi: 10.1145/3477495.3531857. URL https://doi.org/10.1145/3477495.3531857. SPLADE++; arXiv:2205.04733.

Evan Frick, Tianle Li, Connor Chen, Wei-Lin Chiang, Anastasios N. Angelopoulos, Jiantao Jiao, Banghua Zhu, Joseph E. Gonzalez, and Ion Stoica. How to evaluate reward models for RLHF. In *ICLR*, 2025. URL https://arxiv.org/abs/2410.14872. Preference Proxy Evaluations (PPE); arXiv:2410.14872.

Gemma Team. Gemma 4 technical report, 2026. URL https://arxiv.org/abs/2607.02770. Google DeepMind; arXiv:2607.02770.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *COLING*, pp. 6609–6625, 2020. doi: 10.18653/v1/2020.coling-main.580. URL https://aclanthology.org/2020.coling-main.580/.

Yupeng Hou, Junjie Zhang, Zihan Lin, Hongyu Lu, Ruobing Xie, Julian McAuley, and Wayne Xin Zhao. Large language models are zero-shot rankers for recommender systems. In *ECIR*, 2024. URL https://arxiv.org/abs/2305.08845. arXiv:2305.08845.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9. arXiv:2106.09685.

Jerry Huang, Siddarth Madala, Cheng Niu, Julia Hockenmaier, and Tong Zhang. Contextual relevance and adaptive sampling for LLM-based document reranking. In *ACL*, pp. 2076–2089, 2026. doi: 10.18653/v1/2026.acl-long.94. URL https://aclanthology.org/2026.acl-long.94/. TS-SetRank; arXiv:2511.01208.

IBM Granite Team. Granite 4.1 language models, 2026. URL https://huggingface.co/blog/ibm-granite/granite-4-1. Technical blog, released 29 April 2026; dense 3B, 8B and 30B.

Nour Jedidi, Yung-Sung Chuang, James R. Glass, and Jimmy Lin. Don't "overthink" pointwise reranking: Is reasoning truly necessary?, 2025. URL https://arxiv.org/abs/2505.16886. Under review at ICLR 2026; arXiv:2505.16886; OpenReview:kbuGvzBJg4.

Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. LLM-Blender: Ensembling large language models with pairwise ranking and generative fusion. In *ACL*, pp. 14165–14178, 2023. doi: 10.18653/v1/2023.acl-long.792. URL https://aclanthology.org/2023.acl-long.792/.

Anton Korikov, Pan Du, Scott Sanner, and Navid Rekabsaz. Batched self-consistency improves LLM relevance assessment and ranking. In *EMNLP*, pp. 32687–32703, 2025. doi: 10.18653/v1/2025.emnlp-main.1661. URL https://aclanthology.org/2025.emnlp-main.1661/.

Haitao Li, Junjie Chen, Qingyao Ai, Zhumin Chu, Yujia Zhou, Qian Dong, and Yiqun Liu. Calibraeval: Calibrating prediction distribution to mitigate selection bias in LLMs-as-judges. In *ACL*, pp. 16537–16552, 2025. doi: 10.18653/v1/2025.acl-long.808. URL https://aclanthology.org/2025.acl-long.808/.

Xianming Li, Aamir Shakir, Rui Huang, Julius Lipp, Benjamin Clavié, and Jing Li. ProRank: Prompt warmup via reinforcement learning for small language models reranking. In *Findings of ACL*, pp. 1026–1037, 2026. doi: 10.18653/v1/2026.findings-acl.51. URL https://aclanthology.org/2026.findings-acl.51/. released as mxbai-rerank-v2; arXiv:2506.03487.

Xiaobo Liang, Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, and Tie-Yan Liu. R-Drop: Regularized dropout for neural networks. In *NeurIPS*, 2021. URL https://arxiv.org/abs/2106.14448. arXiv:2106.14448.

Chris Yuhao Liu, Liang Zeng, Yuzhen Xiao, Jujie He, Jiacai Liu, Chaojie Wang, Rui Yan, Wei Shen, Fuxiang Zhang, Jiacheng Xu, Yang Liu, and Yahui Zhou. Skywork-Reward-V2: Scaling preference data curation via human-AI synergy. In *ICLR*, 2026. URL https://openreview.net/forum?id=ofgxkMLqic. arXiv:2507.01352.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638. URL https://aclanthology.org/2024.tacl-1.9/.

Yantao Liu, Zijun Yao, Rui Min, Yixin Cao, Lei Hou, and Juanzi Li. RM-Bench: Benchmarking reward models of language models with subtlety and style. In *ICLR*, 2025. URL https://arxiv.org/abs/2410.16184. arXiv:2410.16184.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7. arXiv:1711.05101.

Xuan Lu, Haohang Huang, Rui Meng, Yaohui Jin, Wenjun Zeng, and Xiaoyu Shen. Rethinking reasoning in document ranking: Why chain-of-thought falls short. *arXiv preprint*, 2025. URL https://arxiv.org/abs/2510.08985. arXiv:2510.08985.

Hang Lv, Hongchao Gu, Ruiqing Yang, Liangyue Li, Zulong Chen, Defu Lian, Hao Wang, and Enhong Chen. Learning from emptiness: De-biasing listwise rerankers with content-agnostic probability calibration (CapCal). In *ACL*, pp. 817–826, 2026. doi: 10.18653/v1/2026.acl-short.68. URL https://aclanthology.org/2026.acl-short.68/. arXiv:2604.10150.

Xueguang Ma, Liang Wang, Nan Yang, Furu Wei, and Jimmy Lin. Fine-tuning LLaMA for multi-stage text retrieval (RankLLaMA). In *SIGIR*, 2024. doi: 10.1145/3626772.3657951. URL https://arxiv.org/abs/2310.08319. arXiv:2310.08319.

Saumya Malik, Valentina Pyatkin, Sander Land, Jacob Morrison, Noah A. Smith, Hannaneh Hajishirzi, and Nathan Lambert. RewardBench 2: Advancing reward model evaluation. In *ICLR*, 2026. URL https://openreview.net/forum?id=fb0G86Dewb. arXiv:2506.01937.

Reid McIlroy-Young, Katrina Brown, Conlan Olson, Linjun Zhang, and Cynthia Dwork. Order-independence without fine tuning. In *NeurIPS*, 2024. URL https://arxiv.org/abs/2406.06581. Set-Based Prompting; arXiv:2406.06581.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. MS MARCO: A human generated machine reading comprehension dataset. In *Proceedings of the Workshop on Cognitive Computation: Integrating Neural and Symbolic Approaches 2016*, 2016. URL https://ceur-ws.org/Vol-1773/CoCoNIPS_2016_paper9.pdf.

Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. Document ranking with a pre-trained sequence-to-sequence model (monoT5). In *Findings of EMNLP*, pp. 708–718, 2020. doi: 10.18653/v1/2020.findings-emnlp.63. URL https://aclanthology.org/2020.findings-emnlp.63/.

Justin D. Norman, Michael U. Rivera, and D. Alex Hughes. Reliability without validity: A systematic, large-scale evaluation of LLM-as-a-judge models across agreement, consistency, and bias, 2026. URL https://arxiv.org/abs/2606.19544. arXiv:2606.19544.

Liang Pang, Jun Xu, Qingyao Ai, Yanyan Lan, Xueqi Cheng, and Jirong Wen. Setrank: Learning a permutation-invariant ranking model for information retrieval. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20)*, 2020. doi: 10.1145/3397271.3401104. URL https://arxiv.org/abs/1912.05891. arXiv:1912.05891.

Andrew Parry, Sean MacAvaney, and Debasis Ganguly. Top-down partitioning for efficient list-wise ranking. In *ReNeuIR Workshop at SIGIR*, 2024. URL https://arxiv.org/abs/2405.14589. arXiv:2405.14589.

Ronak Pradeep, Sahel Sharifymoghaddam, and Jimmy Lin. RankZephyr: Effective and robust zero-shot listwise reranking is a breeze! *arXiv preprint*, 2023a. URL https://arxiv.org/abs/2312.02724. arXiv:2312.02724.

Ronak Pradeep, Sahel Sharifymoghaddam, and Jimmy Lin. RankVicuna: Zero-shot listwise document reranking with open-source large language models. *arXiv preprint*, 2023b. URL https://arxiv.org/abs/2309.15088. arXiv:2309.15088.

Jingfen Qiao, Jin Huang, Xinyu Ma, Shuaiqiang Wang, Dawei Yin, Evangelos Kanoulas, and Andrew Yates. LLM-based listwise reranking under the effect of positional bias (DebiasFirst). *arXiv preprint*, 2026. URL https://arxiv.org/abs/2604.03642. arXiv:2604.03642.

Revanth Gangi Reddy, JaeHyeok Doo, Yifei Xu, Md Arafat Sultan, Deevya Swain, Avirup Sil, and Heng Ji. FIRST: Faster improved listwise reranking with single token decoding. In *EMNLP*, pp. 8642–8652, 2024. doi: 10.18653/v1/2024.emnlp-main.491. URL https://aclanthology.org/2024.emnlp-main.491/.

Ruiyang Ren, Yuhao Wang, Kun Zhou, Wayne Xin Zhao, Wenjie Wang, Jing Liu, Ji-Rong Wen, and Tat-Seng Chua. Self-calibrated listwise reranking with large language models. In *Proceedings of the ACM Web Conference 2025 (WWW '25)*, 2025. doi: 10.1145/3696410.3714658. URL https://arxiv.org/abs/2411.04602. arXiv:2411.04602.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, et al. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models (GRPO). *arXiv preprint*, 2024. URL https://arxiv.org/abs/2402.03300. arXiv:2402.03300.

Lin Shi, Chiyu Ma, Wenhua Liang, Xingjian Diao, Weicheng Ma, and Soroush Vosoughi. Judging the judges: A systematic study of position bias in LLM-as-a-judge. In *IJCNLP-AACL*, pp. 292–314, 2025. doi: 10.18653/v1/2025.ijcnlp-long.18. URL https://aclanthology.org/2025.ijcnlp-long.18/. arXiv:2406.07791.

Hongjin Su, Howard Yen, Mengzhou Xia, Weijia Shi, Niklas Muennighoff, Han yu Wang, Haisu Liu, Quan Shi, Zachary S. Siegel, Michael Tang, Ruoxi Sun, Jinsung Yoon, Sercan O. Arik, Danqi Chen, and Tao Yu. BRIGHT: A realistic and challenging benchmark for reasoning-intensive retrieval. In *ICLR*, 2025. URL https://openreview.net/forum?id=ykuc5q381b. arXiv:2407.12883.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. Is ChatGPT good at search? investigating large language models as re-ranking agents (RankGPT). In *EMNLP*, pp. 14918–14937, 2023. doi: 10.18653/v1/2023.emnlp-main.923. URL https://aclanthology.org/2023.emnlp-main.923/.

Raphael Tang, Crystina Zhang, Xueguang Ma, Jimmy Lin, and Ferhan Ture. Found in the middle: Permutation self-consistency improves listwise ranking in large language models. In *NAACL*, pp. 2327–2340, 2024. doi: 10.18653/v1/2024.naacl-long.129. URL https://aclanthology.org/2024.naacl-long.129/.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NeurIPS*, 2017. URL https://arxiv.org/abs/1703.01780. arXiv:1703.01780.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *NeurIPS Datasets and Benchmarks*, 2021. URL https://arxiv.org/abs/2104.08663. arXiv:2104.08663.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022. doi: 10.1162/tacl_a_00475. URL https://aclanthology.org/2022.tacl-1.31/.

Feng Wang, Yuqing Li, and Han Xiao. jina-reranker-v3: Last but not late interaction for listwise document reranking. *arXiv preprint*, 2025a. URL https://arxiv.org/abs/2509.25085. arXiv:2509.25085.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. In *ACL*, pp. 9440–9450, 2024. doi: 10.18653/v1/2024.acl-long.511. URL https://aclanthology.org/2024.acl-long.511/.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *ICLR*, 2023. URL https://arxiv.org/abs/2203.11171. arXiv:2203.11171.

Ziqi Wang, Hanlin Zhang, Xiner Li, Kuan-Hao Huang, Chi Han, Shuiwang Ji, Sham M. Kakade, Hao Peng, and Heng Ji. Eliminating position bias of language models: A mechanistic approach (PINE). In *ICLR*, 2025b. URL https://arxiv.org/abs/2407.01100. arXiv:2407.01100.

William Webber, Alistair Moffat, and Justin Zobel. A similarity measure for indefinite rankings. *ACM Transactions on Information Systems*, 28(4):20:1–20:38, 2010. doi: 10.1145/1852102.1852106. URL https://doi.org/10.1145/1852102.1852106.

Orion Weller, Kathryn Ricci, Eugene Yang, Andrew Yates, Dawn Lawrie, and Benjamin Van Durme. Rank1: Test-time compute for reranking in information retrieval. In *CoLM*, 2025. URL https://arxiv.org/abs/2502.18418. arXiv:2502.18418.

Yanzheng Xiang, Hanqi Yan, Lin Gui, and Yulan He. Addressing order sensitivity of in-context demonstration examples in causal language models (InfoAC). In *Findings of ACL*, pp. 6467–6481, 2024. doi: 10.18653/v1/2024.findings-acl.386. URL https://aclanthology.org/2024.findings-acl.386/.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. C-Pack: Packed resources for general Chinese embeddings. In *SIGIR*, pp. 641–649, 2024. doi: 10.1145/3626772.3657878. URL https://doi.org/10.1145/3626772.3657878. BGE embeddings; arXiv:2309.07597.

An Yang et al. Qwen3 technical report. *arXiv preprint*, 2025. URL https://arxiv.org/abs/2505.09388. Qwen Team; arXiv:2505.09388.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *EMNLP*, pp. 2369–2380, 2018. doi: 10.18653/v1/D18-1259. URL https://aclanthology.org/D18-1259/.

Ziyang Zeng, Dun Zhang, Jiacheng Li, Panxiang Zou, Yudong Zhou, and Yuqing Yang. An empirical study of position bias in modern information retrieval. In *Findings of EMNLP*, pp. 5069–5081, 2025. doi: 10.18653/v1/2025.findings-emnlp.271. URL https://aclanthology.org/2025.findings-emnlp.271/.

Chao Zhang, Shuai Lin, Xuanlin Li, Wenyi Xu, ChengLei Dai, Feiran Zhu, Yi Zhang, and Han Li. CORE: Discovering intrinsic ranking preferences in LLMs via consistent ego-correction, 2026a. URL https://openreview.net/forum?id=ST0Jk1aMsi. OpenReview:ST0Jk1aMsi.

Qianchi Zhang, Hainan Zhang, Liang Pang, Hong-Wei Zheng, and Zhiming Zheng. Stable-RAG: Mitigating retrieval-permutation-induced hallucinations in retrieval-augmented generation. In *ACL*, pp. 25907–25926, 2026b. doi: 10.18653/v1/2026.acl-long.1188. URL https://aclanthology.org/2026.acl-long.1188/.

Jinquan Zheng, Jia Yuan, Jiacheng Yao, Chenyang Gu, Pujun Zheng, and Guoxiu He. Mitigating selection bias in large language models via permutation-aware GRPO (PA-GRPO). In *ACL*, pp. 35125–35143, 2026. doi: 10.18653/v1/2026.acl-long.1621. URL https://aclanthology.org/2026.acl-long.1621/.

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 36 (NeurIPS 2023) Datasets and Benchmarks Track*, 2023. URL https://arxiv.org/abs/2306.05685. arXiv:2306.05685.

Banghua Zhu, Evan Frick, Tianhao Wu, Hanlin Zhu, and Jiantao Jiao. Starling-7B: Improving LLM helpfulness and harmlessness with RLAIF, 2023. URL https://huggingface.co/datasets/berkeley-nest/Nectar. releases the Nectar ranking dataset.

The appendix is organized as follows, each answering a different question:

App. A the scope and limitations of the study.
App. B how a score is read from the prompt, and what reproduces a run.
App. C the origin of the order effect, and controls that exclude other explanations.
App. D whether cheaper alternatives suffice, what they cost, and first-stage robustness.
App. E whether the effect is specific to reranking, on QA and on response ranking.
App. F four consumers of a ranking, and what training changes for each.
App. G how many candidates to serve, and what width costs in quality and latency.
App. H per-collection results, alternative first stages, and external rerankers.

Unless a caption says otherwise, every table and figure below uses Qwen3-4B at $B = 20$ on the 18 reranking collections at seed 42, over $M = 10$ random permutations. A basis line states where a table or figure departs from that.

A Scope and limitations

The serving pass takes its order from the first stage, so the permutations we measure under are not the ones it runs on. Controls on the first-stage order leave both τ-PSI and the output's correlation with BM25 nearly unchanged, so the measurement is representative of the served condition (Table 6). They are also uniform draws, so the spreads we report describe typical rather than worst-case presentations. The threshold, the reader and the preference model are frozen rather than refitted to each scorer's output. A co-adapted consumer could absorb some of the instability we measure, so the reductions we report may be smaller in a fully retrained pipeline. We measure the chosen/rejected pairs supplied to preference training, not the effect of fitting a downstream model to those labels; whether the instability survives downstream optimization remains untested.

Round-robin partitioning and logit calibration are two prompt-time constructions, not all possible ones, and both run under the single answer skeleton of App. B.1.1. InvariRank (Bito et al., 2026) restricts cross-candidate attention to obtain one-pass order invariance; App. D.2 ports the same mechanism from Wang et al. (2025b) and finds it costs both quality and stability at this readout, so

built-in invariance is unavailable here rather than untried. PA-GRPO (Zheng et al., 2026) rewards consistent decisions with reinforcement learning, and its reward is defined on the decision rather than on a score. Redefining it over the continuous score our objective targets would change the method rather than merely porting it.

The objective targets stability and the part of quality that order-averaging recovers, not raw quality: a pointwise $B = 1$ scorer trained in our own pipeline matches our students on nDCG@10. Heavily supervised commercial rerankers exceed them, on far larger relevance supervision than ours. The case for scoring candidates together ($B > 1$) does not cover pools of two: no collection we evaluate has fewer than four candidates (Table 3). The one recurring collection-level cost is ArguAna, the suite's only counterargument-retrieval collection (Table 35). Rescoring one permutation does not return identical grades for any scorer here: ours reproduce 0.985 of the retained set at fixed order, and GPT-5.4's floor is unmeasured (Table 26). Its overlap is therefore lowered by noise as well as by order, which flatters our comparison. Nearly half its candidate pairs tie, so its reported figures turn on the tie-break: it reaches 0.468 nDCG@10 under the order-independent rule we keep against 0.511 under input-order ties. That rule also gives it a lower pair-flip rate than OC-SFT, which reverses under the common stable sort. The τ-PSI we withhold would favour it too: at 0.118 it would rank second in the reranking column (Table 1), and about sixth once tied pairs are discounted. The convention we report is the one that favours the baseline.

Finally, the scorer uses no chain-of-thought (Lu et al., 2025; Jedidi et al., 2025; Weller et al., 2025), so reasoning-intensive retrieval (BRIGHT, Su et al., 2025) and reasoning rerankers in their own regime are untested. All collections are English, and long-context corpora that overflow the $B = 20$ prompt fall outside the geometry studied here (App. B.1).

B Setup: readout, evaluation, and configuration

We specify how a score is read out of a single prompt, including the placeholder and answer-skeleton choices it depends on (App. B.1), which collections carry each measurement (App. B.2), the consolidated training and evaluation configuration (App. B.3), and the interval and bolding protocol for Table 1 (App. B.4).

B.1 Score readout and prompt

The reranker scores a query against B candidate documents in one prompt. After a short grading instruction, the prompt places the query and the $B = 20$ documents, each tagged `[i]`, then appends a fixed `Grades:` skeleton carrying one placeholder grade per slot. Documents are truncated to 1200 characters on TREC Deep Learning and non-reranking corpora, and 500 on the BEIR and legal collections. We score the fixed `Grades:` skeleton in one forward pass, reading four grade-token probabilities per candidate rather than generating grades (§4). No token is sampled, so a score is determined by the weights and the prompt, and no decoding temperature or generation seed enters. Candidate order is the variation we report; with order held fixed, engine-call batching still moves retained-set overlap, but far less than reordering does (Table 26). GPT-5.4 is the one exception, and its decoding parameters are not configurable. The scoring client's nominal temperature (1.0) and top-p (10^{-8}) do not reach the provider. The interface omits temperature, the client does not forward top-p, and no generation seed is exposed; the run seeds shuffle documents rather than generation. FIRST (Reddy et al., 2024) also uses logits instead of decoding, but takes a listwise ranking from the first generated identifier, whereas we take one score per candidate. The expectation also keeps scores near-continuous, so ties rarely decide the ranking. At $B = 20$, 99.6% of scores within a query and permutation are distinct, and re-ranking the saved scores under an order-independent tie-break on document id moves every τ-PSI value we report by less than 0.0005. The full prompt is shown below.

```
System: You are a search relevance grader. For each numbered document,
    output one fixed
        integer relevance grade in {0, 1, 2, 3}. Do not rank the
    documents and do not explain.

User:   Instruction: Given a web search query, retrieve relevant passages
     that answer the query
        Query: <q>
```

```
        Documents:
        [1] <doc 1 text>            # whitespace-normalized + truncated
    to 1200 chars
         ...
        [20] <doc 20 text>

        Evaluate every document independently for relevance to the query.
        Assign exactly one integer relevance grade to every document:
        - 3 = directly and completely answers the query
        - 2 = strongly relevant but incomplete
        - 1 = weakly relevant or topical background
        - 0 = not relevant
        Output one line per document in the form: [<id>] Grade: <0|1|2|3>

Assistant:  Grades:                 # not sampled; a fixed skeleton is
    scored
        [1] Grade: 0                #   read P(0),P(1),P(2),P(3) at this
    slot
         ...                        #   score = E[g]/3
        [20] Grade: 0
```

The readout and the placeholder are both free choices, so we selected each once on the off-the-shelf Qwen3-4B scorer and held them fixed thereafter. Table 2 reports both comparisons, comparing the argmax grade, $P(g \geq 2)$ and the expected grade, all three read off the same saved grade distribution. The argmax grade is lower than either probability readout on the aggregate grid. The two probability readouts are nearly level; $P(g \geq 2)$ leads by less than 0.001 nDCG@10, so we retain the expected grade as the graded readout and hold it fixed.

Table 2: **Score-readout and answer-skeleton placeholder choices.** Basis: off-the-shelf Qwen3-4B over the 18 reranking collections. Quality is one-permutation nDCG@10; instability is τ-PSI at $B = 20$ over $M = 10$ random permutations. Higher quality and lower instability are better; the final two columns count collections with lower / higher instability than `Grade: 0`. Bold marks the selected protocol choice rather than the per-column optimum.

	Quality, one permutation		Order instability ($M = 10$)		
Choice	nDCG@10 ↑	Wins	τ-PSI ↓	Lower	Higher
Score readout, quality only					
argmax grade	0.3365	1 / 18	—	—	—
$P(g \geq 2)$	0.3701	9 / 18	—	—	—
expected grade	**0.3694**	**8 / 18**	—	—	—
Answer skeleton placeholder					
`Grade: 0`	**0.3694**	**17 / 18**	**0.2983**	ref.	ref.
`Grade: 1`	0.3541	0 / 18	0.2928	11	7
`Grade: 2`	0.3413	1 / 18	0.3348	6	12
`Grade: 3`	0.2013	0 / 18	0.3857	0	18

B.1.1 Answer-skeleton placeholder

Varying only the in-format placeholder from `Grade: 0` through `Grade: 3`, quality decreases monotonically from the selected `Grade: 0` (Table 2). `Grade: 1` has the lowest instability point estimate but lower quality, while `Grade: 2` and `Grade: 3` worsen both axes. No in-format placeholder improves both quality and instability over the selected `Grade: 0`, so placeholder choice is not a substitute for training.

B.2 Evaluation collections

The eleven public out-domain collections are the BEIR subset introduced for listwise LLM reranking by Sun et al. (2023) and carried forward by the distilled listwise rerankers that followed (Pradeep et al.,

2023a;b). That subset includes the three collections later work often drops: Signal-1M, TREC-NEWS and Robust04. Those papers report DL19 and DL20 on MS MARCO v1; we add the three TREC DL v2 tracks, DL21 through DL23, so the sixteen extend that suite rather than merely reuse it. Legal-A and Legal-B are proprietary collections from a commercial legal search provider, holding professionally authored legal content with relevance judgments made by attorneys. Their candidate lists come with the internal fixtures rather than being retrieved by BM25.

Table 3: **Evaluation collections across the three tasks.** Counts are evaluated queries or prompts after filtering; protocols are in §5.

Task / group	Collection	Domain / role	Queries / prompts	Candidate pool / protocol
Reranking: in-domain	DL19	web passage; TREC DL	43	BM25 top 100; graded judgments
	DL20	web passage; TREC DL	54	BM25 top 100; graded judgments
	DL21	web passage; TREC DL v2	53	BM25 top 100; graded judgments
	DL22	web passage; TREC DL v2	76	BM25 top 100; graded judgments
	DL23	web passage; TREC DL v2	82	BM25 top 100; graded judgments
Reranking: public OOD	NFCorpus	biomedical	308	BM25 top 100
	FiQA	financial QA	648	BM25 top 100
	Touché-2020	argument retrieval	49	BM25 top 100
	ArguAna	counterargument retrieval	1,406	BM25 top 100
	Climate-FEVER	climate fact-checking	1,535 (1,381 with verdict labels)	BM25 top 100
	TREC-COVID	biomedical	50	BM25 top 100
	DBPedia	entity retrieval	400	BM25 top 100
	SciFact	scientific fact-checking	300	BM25 top 100
	Signal-1M	tweet / news	97	BM25 top 100
	TREC-NEWS	news	57	BM25 top 100
	Robust04	news	249	BM25 top 100
Reranking: legal	Legal-A	legal	97	precomputed top 100
	Legal-B	legal	175	precomputed top 100
Multi-doc QA	HotpotQA	multi-hop QA; in-domain	200	10 passages; 2 supporting
	2WikiMultiHopQA	multi-hop QA; zero-shot	200	10 passages; 2 supporting
	MuSiQue	multi-hop QA; zero-shot	200	10 passages; 2 supporting
Response ranking	RewardBench-2	response quality	1,763	4 responses
	Nectar	response preference	434	7 responses; 64 overlaps removed
	PPE-MATH	math correctness	512	8 responses
	PPE MMLU-Pro	knowledge correctness	512	8 responses
	RM-Bench	adversarial preference	1,327	6 responses

B.3 Consolidated configuration

Every task selects its consistency weight λ and its checkpoint on that training task's own held-out split, using held-out ranking quality as the only criterion. The grid is $\lambda \in \{0.5, 1, 2, 3, 4, 5\}$. The grid's upper end is selected for passage reranking; the same rule selects interior values elsewhere, λ=3 for multi-document question answering and λ=1 for response ranking (Table 4), so the selection is not saturating by construction. We did not train above λ=5. OC-SFT is the only variant with a consistency weight to select. The alternative regularizers of App. D.2, permutation augmentation and DebiasFirst use fixed recipes and are reported at the checkpoint that maximizes the same held-out ranking-quality criterion.

Table 4 gives what differs between the three tasks. The settings they share:

- LoRA (Hu et al., 2022): rank 16, alpha 32, dropout 0.05, on the seven attention and MLP projections (`q`, `k`, `v`, `o`, `gate`, `up`, `down`) in every transformer layer.
- Optimizer: AdamW (Loshchilov & Hutter, 2019) ($\beta_1 = 0.9$, $\beta_2 = 0.95$, weight decay 0.01), learning rate 2e-4 with a cosine schedule and a 10% linear warmup, 1 epoch; per-device batch 1 with gradient accumulation 8 across the 8-GPU node, an effective batch of 64; bf16 with FlashAttention-2.

- OC-SFT only: the two views of a window are permutations drawn from seeds derived from the query and window identifiers. At one epoch, each window has a fixed pair of orders, and the anchored view is the first of the two rather than the first stage's own order. The loss uses the raw 0–3 grade scale; the $[0, 1]$ normalization is applied at scoring time, so λ is on the grade scale.
- OC-SFT only: λ ramps linearly from 0 to its target over the first 500 optimizer steps, separately from the learning-rate warmup.
- Training hardware: 8-GPU DDP nodes, with an A10G for the 1.7B student, A100-40GB for the 4B to 8B Qwen3 students and the small Gemma and Granite students, and A100-80GB for the 14B, 32B and 30B-class students.
- Evaluation hardware: a L40S-48GB GPU for evaluation, scoring and for the τ-PSI computation.

Table 4: **What differs between the three tasks.** Every other setting is shared.

	Passage reranking	Multi-document QA	Response ranking
Training data	MS MARCO, ∼30K queries and ∼3M document labels	HotpotQA distractor, 30K questions	UltraFeedback, ∼30.7K prompts
Training width	$B = 20$	$B = 10$	$B = 4$
Rubric	graded 0–3 topical relevance	answer support, binary	response quality
Consistency weight	$\lambda = 5$	$\lambda = 3$	$\lambda = 1$
Selection split	held-out MS MARCO	disjoint 500 questions	held-out UltraFeedback
Quality metric	nDCG@10	nDCG@10	nDCG@1

B.4 Inference for Table 1

Intervals follow §5's protocol at R=20,000 resamples; seeds are averaged rather than resampled, since three cannot support a sampling model over training randomness. Holm adjustment covers every contrast between OC-SFT and the other trained variants and each has adjusted $p \leq 0.0013$ with all seeds favorable.

C Mechanism and stability controls

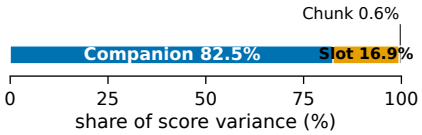
We decompose order dependence into slot and window-index effects plus a residual consistent with companion coupling, test transfer to pool-membership changes and marker renaming (App. C.1), rule out measurement artifacts (App. C.2), and test robustness across metrics (App. C.3).

C.1 Mechanisms and perturbation transfer

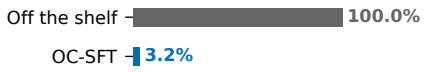
In passage reranking, the first stage supplies the top 100 candidates, which are scored in five $B = 20$ windows, so a document's grade can vary with its slot, window index and window companions (Figure 5). For each document, we subtract its mean score across presentations, removing its fixed score level. The decomposition assigns separate shares of the remaining within-document variance to slot and window-index fixed effects; the residual, which is consistent with companion coupling but may include unmodelled interactions, accounts for more than four fifths of off-the-shelf per-document score variance. OC-SFT removes 96.8% of that residual variance.

Averaged over queries, each batch position p has a mean order residual $\psi(p)$; under the cumulative `Grade: 0` skeleton of App. B.1, the off-the-shelf profile is monotone and early-over-scored. This differs from the U-shaped "lost in the middle" effect in long-context QA and resembles the primacy pattern attributed to causal attention (Liu et al., 2024; Tang et al., 2024; Parry et al., 2024). OC-SFT flattens both the pooled and gold-only profiles. An alternate in-format placeholder from the set compared in App. B.1 attenuates the off-the-shelf profile, so part of the slot effect belongs to the readout. A forced-position nDCG control shows neither monotone primacy nor a U-shape on any collection, so the score-level slot residual does not become a ranking-level positional preference (Table 6).

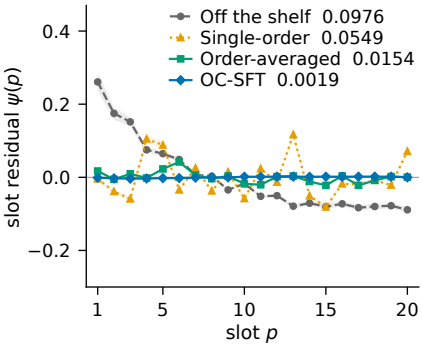
Order robustness transfers to unseen pool-membership changes (Table 5). During training, every shuffled view contains the same candidate documents. The loss never observes a candidate being replaced or removed. To test transfer beyond that perturbation, we sample one document from the top 100 uniformly with a stable per-query seed and either replace it with the next first-stage



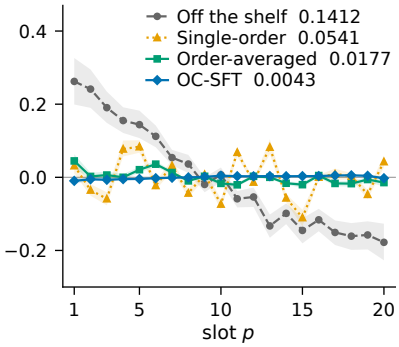
(a) Off-the-shelf within-document score-variance decomposition over 11 reranking collections.



(b) Residual variance after slot and window-index controls, relative to off the shelf over four matched collec-



(c) Mean slot residual $\psi(p)$ over all documents; RMS is in the legend.



(d) Mean slot residual over gold-positive documents on the same scale.

Figure 5: **Residual variation consistent with companion coupling dominates off-the-shelf score variance; OC-SFT removes nearly all of it and flattens the slot residual.** Basis: Qwen3-4B; (a–b) use eleven and four collections, while (c–d) use a separate held-out MS MARCO sample of 200 queries at 500 permutations.

Table 5: **Pool perturbations change decisions without changing mean quality.** Lower is better except for signed mean ΔnDCG@10. Basis: Qwen3-4B, seed 42, 15 reranking collections, up to 100 judged queries each; rankings are restricted to the 99 documents shared by the two pools; the sampled rank, dropped document and replacement document are identical across the four columns for every query. DL21–23 are excluded because their first-stage files stop at rank 100 and do not contain the replacement candidate at rank 101.

Quantity	Off the shelf	Single-order	Order-avg.	OC-SFT
Pool-PSI	0.102	0.099	0.063	**0.031**
Top-10 changed	74.6%	65.0%	51.5%	**36.3%**
Mean ΔnDCG@10	−0.001	+0.001	+0.001	+0.001
Mean absolute ΔnDCG@10	0.0356	0.0240	0.0153	**0.0113**

candidate, appended at the end of the pool, or drop it, hold the serving order fixed and compare only the shared documents. Pool-PSI applies the τ-PSI transform $(1-\tau)/2$ to the original and perturbed rankings over those shared documents. On all 15 collections, order-averaged distillation is between single-order distillation and OC-SFT on Pool-PSI, and the same aggregate ordering holds for top-10 churn; mean nDCG@10 remains level. Averaging the teacher targets therefore provides part of the pool robustness, and the consistency penalty provides the remaining reduction. The drop-only control changes Pool-PSI by less than 0.001 for any variant, so the reduction reflects changed companions rather than the introduction of rank 101.

Transfer depends on the perturbation: order-consistency training generalizes to changed pool membership but not to marker renaming (Tables 5 and 6). Candidate markers are arbitrary identifiers attached to the documents, so renaming them cannot change relevance but does move off-the-shelf scores. Neither order-consistency training nor exposure to varied markers during SFT removes this sensitivity. Adding a marker-consistency penalty, the same term with marker renamings substituted for order shuffles, lowers held-out marker τ-PSI on all six held-out cells.

The first-stage deficit follows window assignment rather than slot direction or copying (Tables 1 and 6). Round-robin assignment improves the untrained scorer by distributing adjacent first-stage

ranks across windows (Table 1). Reversing the candidate order does not repair the deficit: quality stays below both the forward and random-permutation rankings, and τ-PSI is effectively unchanged. Because BM25 supplies the first stage, the forward condition presents candidates in BM25 order; its output ranking is only marginally more correlated with BM25 than the output from a random permutation. Together, these controls attribute the deficit to which candidates share a window.

Table 6: **Three controls separate order mechanisms and perturbation transfer.** Block (a) tests whether the slot residual reaches ranking quality; block (b) tests transfer to marker renaming; block (c) tests the first-stage-order deficit. Basis: blocks (a) and (c) use Qwen3-4B on the 18 reranking collections; block (b) uses Qwen3-1.7B and Qwen3-4B on DL19, NFCorpus, and Legal-A , with Roman-numeral and random three-character marker schemes held out from training.

(a) Ranking-level slot preference		
Diagnostic	Collections with primacy	Collections with U-shape
Forced-position nDCG@10	0/18	0/18

(b) Marker-renaming transfer			
Measure	Minimum	Maximum	Coverage / wins
Held-out marker τ-PSI	0.035	0.059	6 cells
Penalty improvement over marker-varied SFT	0.013	0.037	6/6

(c) First-stage-order controls			
Measure	Reverse	Forward	Random permutation
nDCG@10	0.347	0.370	0.415
τ-PSI	0.2992	0.2983	—
Kendall correlation with BM25	—	0.166	0.158

C.2 Validity controls for stability

Table 7: **Validity controls for the stability result.** Block (a) separates stability from quality gain; block (b) tests score collapse and tied rankings; block (c) tests movement at the ranking head. Blocks (b–c) use the 18 reranking collections. Basis: Qwen3-4B, seed 42, $M = 10$ random permutations for blocks (b–c), and the document-ID tie-break.

(a) Quality controls				
Diagnostic	Collection	Single-order	Order-avg.	OC-SFT
$\lvert\mathrm{corr}(\Delta\mathrm{nDCG@10},$ off-the-shelf $\tau\text{-PSI})\rvert$	DL19	0.257	0.194	0.196
	DL20	0.070	0.037	0.106
Queries improving over off the shelf	DL19	93.0%	97.7%	90.7%
	DL20	90.7%	90.7%	88.9%
Regressions relative to single-order	DL19	reference	37.2%	27.9%
	DL20	reference	29.6%	24.1%

(b) Score-geometry controls			
Diagnostic	Single-order	Order-avg.	OC-SFT
Residual normalized by score dispersion	0.328	0.237	0.154
Exact tied pairs	0.017%	0.050%	0.351%

(c1) Ranking-head levels	
Variant	Top-10 instability
Single-order	0.343
Order-avg.	0.275
OC-SFT	0.193

(c2) Ranking-head paired contrasts			
Contrast	Estimate	95% CI	Collection wins
Single-order − OC	+0.150	[+0.140, +0.159]	OC 18/18
Order-avg. − OC	+0.081	[+0.074, +0.087]	OC 18/18
Single-order − Order-avg.	+0.069	[+0.061, +0.076]	Order-avg. 18/18

Lower τ-PSI reflects smaller permutation-induced movement rather than improved relevance, tied scores or tail-only changes (Table 7). Per-query quality gain against the off-the-shelf scorer is weakly related to starting instability and is non-negative for 89–98% of queries across the three trained variants on DL19 and DL20. Quality gains are aggregate: both order-averaged distillation and OC-SFT regress on some queries relative to single-order distillation. Scale-normalized residuals fall monotonically across the three variants while ties remain rare, so score collapse does not account for lower instability. Top-10 instability is $1 - \mathrm{RBO@10}$ at persistence p=0.9 (Webber et al., 2010), and it decreases in the same order on every collection. The reduction includes the highest-ranked documents, including the top five passages passed to the downstream reader (§5).

We do not report τ-PSI for GPT-5.4, whose grades do not reproduce at a fixed order (App. B.1). Rescoring the same permutation, on one query from each of DL19, FiQA and Climate-FEVER, agrees on 0.963 of grades and leaves a same-order τ-PSI floor of 0.023. Sampling no token removes that source for the open-weight scorers, but not the effect: at fixed candidate order they still reproduce 0.985 of the retained set (Table 26).

C.3 Robustness across metrics

OC-SFT remains best across alternative quality and stability metrics (Table 8). The four columns test average quality, worst-permutation quality, dispersion across permutations and rank instability, so the ordering does not depend on τ-PSI alone. From the off-the-shelf scorer to OC-SFT, worst-permutation nDCG improves more than mean-permutation nDCG; because order-spread is mean minus worst, its reduction equals the difference between those gains.

Table 8: **The variant ordering holds across metrics.** Columns report mean and worst nDCG, order-spread and τ-PSI for four variants; higher nDCG and lower instability are better. Basis: eleven dense bases by the 18 reranking collections and $M = 10$ random permutations; nDCG uses the order-independent document-ID tie-break defined in App. B.1; order-spread is mean minus worst before rounding.

Variant	mean-perm nDCG	worst-perm nDCG	order-spread	τ-PSI
Off the shelf	0.4208	0.2940	0.1268	0.2874
Single-order	0.4606	0.3736	0.0870	0.2245
Order-avg.	0.4726	0.4076	0.0650	0.1388
OC-SFT	**0.4771**	**0.4296**	**0.0475**	**0.0858**

D Mitigation, supervision cost, and robustness

We re-measure the trained gain with the partition held fixed (App. D.1) and compare OC-SFT with mitigation baselines and architectural controls (App. D.2). We then price its supervision and training cost (App. D.3), weigh inference-time averaging against training for consistency (App. D.4), measure sample efficiency (App. D.5), repeat the averaging comparison on consumers (App. D.6), and check that the gain survives a different first stage (App. D.7).

D.1 Partition controls and matched training gain

Round-robin partitioning does not explain the OC-SFT gain (Table 9). Round-robin raises nDCG@10 for both scorers but raises the off-the-shelf scorer more, reducing the training gain without eliminating it. Across the six listed variants, the partition change moves τ-PSI by at most 0.0053. For the off-the-shelf scorer, neither retained-set overlap, verdict flip on either collection, nor pair flip changes measurably, with all four intervals covering zero.

Table 9: **Round-robin partitioning does not explain the trained gain.** Block (a) gives first-stage-order single-pass nDCG@10 under each partition rule; training gain is OC-SFT minus off the shelf. Block (b) gives round-robin minus contiguous τ-PSI over $M = 10$ random permutations. Block (c) gives round-robin minus contiguous consumer output for the off-the-shelf scorer; every interval covers zero. Basis: Qwen3-4B and 18 reranking collections; blocks (a–b) use seed 42 and equal-weight collection means, while block (c) uses $M = 10$ permutations under each rule and the two claim-verification collections for verdict flip. Block (a) intervals, and block (c)'s retained-set-overlap and selected-pair rows, are paired collection bootstraps with 20,000 resamples; the two verdict-flip rows are paired query bootstraps within SciFact and Climate-FEVER.

(a) Training gain under fixed partitioning

Partition rule	Off the shelf	OC-SFT	Training gain [95% CI]
Contiguous	0.370	0.459	+0.089 [+0.069, +0.110]
Round-robin	0.422	0.474	+0.052 [+0.034, +0.073]
Contiguous minus round-robin	−0.052	−0.016	+0.036 [+0.022, +0.050]

(b) Partition effect on instability

Variant	Round-robin minus contiguous τ-PSI
Off the shelf	−0.0010
Single-order	+0.0053
Order-avg.	+0.0029
Perm. aug.	+0.0005
DebiasFirst	+0.0008
OC-SFT	−0.0004

(c) Off-the-shelf consumer effect

Consumer	Change	95% CI
Retained-set overlap	+0.004	[−0.007, +0.016]
Verdict flip, SciFact	+0.0147	[−0.0102, +0.0404]
Verdict flip, Climate-FEVER	−0.0002	[−0.0062, +0.0058]
Selected pair	+0.0004	[−0.0033, +0.0034]

D.2 Mitigation baselines and architectural controls

No mitigation baseline reaches OC-SFT's mean τ-PSI over the 18 reranking collections (Table 10). EMA closes 71% of the stability difference between single-order distillation and OC-SFT; order-averaged distillation at T=10, permutation augmentation, and DebiasFirst close 63–64%. Averaging fewer teacher orders does not reach that: the τ-PSI floor over T=2, 3 and 5 is 0.153, and the curve is not saturated at five, so T=10 is not an over-provisioned comparator. LoRA dropout leaves τ-PSI level with single-order distillation, KL regularization worsens it, and CapCal remains at the off-the-shelf level (Table 1). DebiasFirst and permutation augmentation differ only in propensity weighting and yield 0.128 and 0.129 τ-PSI, so the weighting does not measurably help in our score-regression port.

PINE's patched attention breaks our multi-candidate slot-grade readout (Wang et al., 2025b). On DL19 at $B = 20$, the ported scorer loses 0.122 nDCG@10 and adds 0.114 τ-PSI relative to the base Qwen3-4B scorer. A segmented prompt recovers most of the nDCG loss, and scoring one document at a time matches the base, so the failure arises from attention across documents rather than the readout alone.

Table 10: **Mitigation baselines and regularizers on the Qwen3-4B student.** Higher nDCG@10 and lower τ-PSI are better. Each variant is the mean $\pm$ sample standard deviation. Bold is a descriptive seed-spread heuristic, not the Holm-adjusted marking of Table 1: a lead larger than the seed spread of the rows it leads. The τ-PSI column clears that heuristic, whereas the quality column does not, so quality is unmarked. % closed is the fraction of the stability difference between single-order distillation and OC-SFT that the method removes. T is the number of shuffled teacher orders averaged into each label; the T=2, 3 and 5 draws are the first T orders of the T=10 pool rather than independent draws. Basis: Qwen3-4B, 18 reranking collection means, 3 training seeds.

Variant	Mean nDCG@10	Mean τ-PSI	% closed
Single-order (T=1)	0.449 ± 0.003	0.209 ± 0.008	0%
+ LoRA dropout 0.20	0.448 ± 0.004	0.208 ± 0.001	+0.8%
+ KL-to-base $\beta = 0.05$	0.447 ± 0.002	0.231 ± 0.012	−17.5%
+ EMA mean-teacher $\alpha = 0.5$	0.453 ± 0.003	0.119 ± 0.003	71.4%
Order-avg. T=2	0.451 ± 0.002	0.176 ± 0.008	26.2%
Order-avg. T=3	0.451 ± 0.000	0.153 ± 0.002	44.4%
Order-avg. T=5	0.450 ± 0.001	0.161 ± 0.009	38.1%
Order-avg. T=10	0.455 ± 0.002	0.130 ± 0.007	62.7%
Perm. aug.	0.455 ± 0.002	0.129 ± 0.010	63.5%
DebiasFirst	0.454 ± 0.001	0.128 ± 0.005	64.3%
OC-SFT ($\lambda^* = 5$)	0.459 ± 0.002	$\mathbf{0.083 \pm 0.002}$	100%

D.3 Supervision and training cost

To separate information in the labels from exposure during student training, we cross teacher scoring width ($B = 1$ or 20) with student training width ($B = 1$ or 20). All four variants use Qwen3-4B as teacher and student; App. H.4 varies teacher identity separately.

Batched-teacher labels improve student nDCG@10 at both training widths, although the gain is smaller when the student also trains at $B = 20$ (Table 11). The pointwise teacher ranks known MS MARCO positives higher under direct teacher metrics, yet its labels train a worse student. Remapping the pointwise labels to the batched teacher's score distribution while preserving document order does not close the student gap. Neither teacher-side ranking quality nor marginal score distribution therefore explains the benefit of batched-teacher labels.

Table 11: **Batched-teacher labels improve student nDCG@10 at both training widths.** The upper block reports the difference between students trained on batched- and pointwise-teacher labels; positive values favor batched-teacher labels. All students are evaluated at serving width $B = 1$ in first-stage order. Rank-quantile matching assigns the batched teacher's sorted score values to documents in the pointwise teacher's rank order. The residual is the batched-label student's nDCG@10 minus that of the recalibrated pointwise-label student; minimum and maximum are over three evaluated checkpoints. Basis: mean over the 18 reranking collections; Qwen3-4B; training seed 42; 100,000 paired collection-bootstrap samples.

Student training width	ΔnDCG@10	95% CI	Wins (of 18)
$B_{\text{train}} = 1$	+0.0291	[+0.0195, +0.0385]	17
$B_{\text{train}} = 20$	+0.0097	[+0.0023, +0.0169]	13

Control	Minimum residual	Maximum residual
Rank-quantile matching	+0.0307	+0.0333

Increasing student views beyond $N = 2$ does not help, while additional teacher passes mainly improve stability in a plain-SFT control (Table 12). From N=2–8, nDCG@10 and τ-PSI remain within the ranges in Table 12 while forward-pass training cost rises fourfold. On a plain-SFT student with the teacher held fixed, increasing teacher permutations from T=5 to T=40 changes nDCG@10 by +0.0042 and τ-PSI by −0.0398 at 8× the label-generation cost.

Table 12: **Extra student views raise training cost; extra teacher passes mainly lower τ-PSI.** N is the number of shuffled student views per training step; T is the number of teacher permutations averaged into each stored label. Student-view rows report the range over $N \in \{2, 3, 4, 8\}$; teacher-label rows report the change from T=5 to T=40. Cost is relative to N=2 training for student-view rows and T=5 labeling for teacher-label rows. Basis: Qwen3-4B at the selected operating point; student-view rows are a five-collection mean over DL19, DL20, Touché-2020, NFCorpus and Legal-A at a single seed, not the 18-collection default, so they are not comparable to the means in Table 1.

Stage varied	Sweep	Metric	Observed range / change	Forward-pass cost
Student training	$N \in \{2, 3, 4, 8\}$	nDCG@10	range 0.4786–0.4795	100–400% training
Student training	$N \in \{2, 3, 4, 8\}$	τ-PSI	range 0.067–0.077	100–400% training
Averaged silver, plain SFT	$T = 5 \rightarrow 40$	nDCG@10	change $+0.0042$	800% labeling
Averaged silver, plain SFT	$T = 5 \rightarrow 40$	τ-PSI	change -0.0398	800% labeling

Re-distillation helps in the first round and then largely saturates (Table 13). Qwen3-4B gains mostly in stability on the first round and changes by only 0.0011 on each metric in the second, with τ-PSI increasing; Gemma-E4B retains a smaller second-round gain. A second round therefore buys a change of that size for another full labeling and training cycle, so iterating further is not an efficient way to add stability.

Table 13: **Iterative self-distillation gains diminish after the first round.** Positive ΔnDCG@10 and negative $\Delta\tau$-PSI are improvements. Round 1 is measured against the initial trained student and round 2 against round 1. Basis: 18 reranking collections, one iterative training lineage per base.

Base	Round	ΔnDCG@10	$\Delta\tau$-PSI
Qwen3-4B	1	$+0.0069$	-0.0183
Qwen3-4B	2	$+0.0011$	$+0.0011$
Gemma-E4B	1	$+0.0141$	-0.0311
Gemma-E4B	2	$+0.0036$	-0.0077

D.4 Inference-time amortization

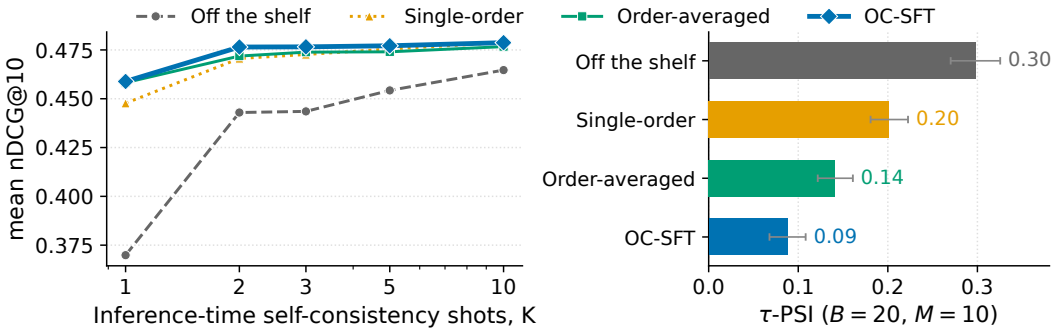
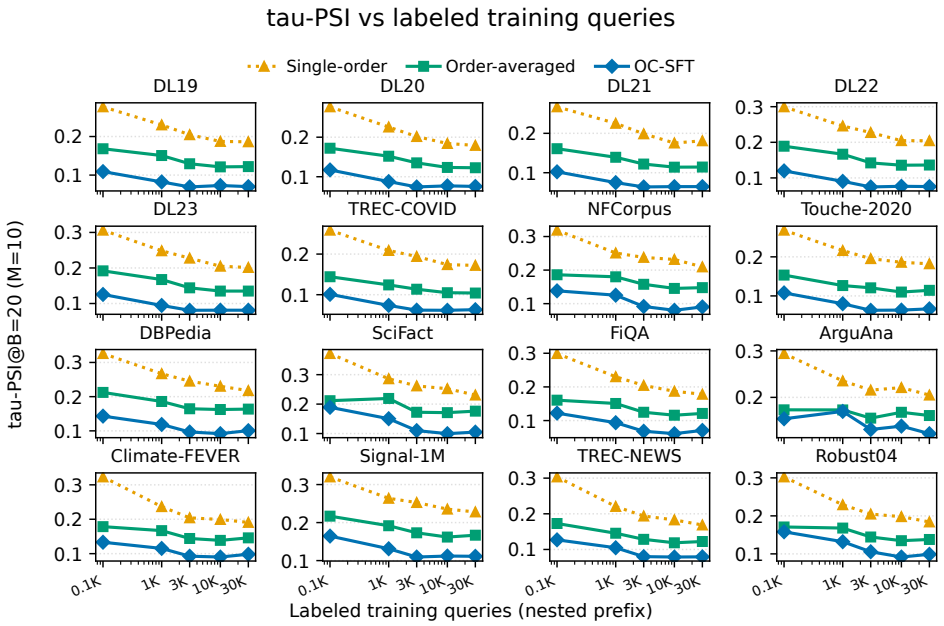


Figure 6: **OC-SFT at $K = 1$ is within 0.006 nDCG@10 of the off-the-shelf $K = 10$ ensemble.** Left: nDCG@10 against self-consistency permutations K, higher is better. Right: single-pass τ-PSI at $B = 20$ for off the shelf, single-order distillation, order-averaged distillation and OC-SFT, evaluated over $M = 10$ random permutations; lower is better.

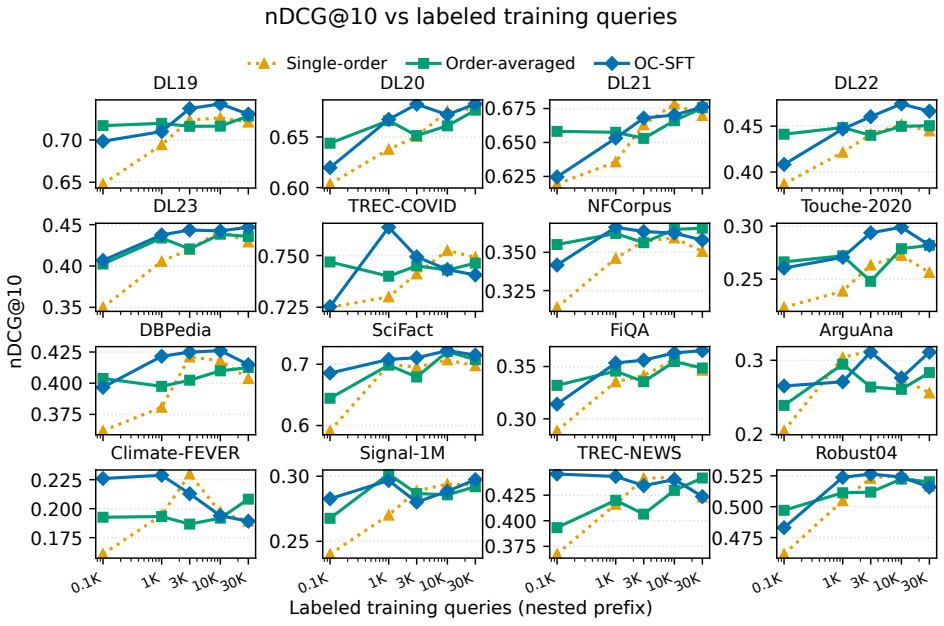
Averaging raises nDCG@10, while training supplies most of the τ-PSI reduction (Figure 6). At $K = 5$, the off-the-shelf scorer still trails its $K = 10$ nDCG@10, so we retain $K = 10$ as the quality comparator. OC-SFT has the lowest single-pass τ-PSI among the four variants. From $K = 1$ to $K = 10$, averaging raises off-the-shelf nDCG@10 by 0.095 but each trained variant by at most 0.035, so training has already captured most of the nDCG@10 gain that permutation averaging provides.

D.5 SAMPLE EFFICIENCY

τ-PSI saturates with fewer labeled queries than nDCG@10 (Figure 7). With 1K labeled queries, OC-SFT is more stable than order-averaged distillation trained on 30K queries on every collection except ArguAna. For quality, models trained from one teacher order need about 3K queries to approach their 30K nDCG@10, whereas models trained from ten-order averages saturate earlier.



(a) τ-PSI vs. labeled queries



(b) nDCG@10 vs. labeled queries

Figure 7: **Stability saturates with fewer labeled queries than quality.** (a) Student τ-PSI, lower is better; (b) nDCG@10, higher is better, both against the number of unique labeled training queries over nested pools. Basis: Qwen3-4B; 16 public reranking collections; Legal-A and Legal-B follow the same pattern.

D.6 SERVING-TIME AVERAGING AFTER TRAINING

App. D.4 sweeps the serving budget for nDCG@10 and τ-PSI. We repeat that sweep on the three consumers, so that whether training removes what averaging buys is answered on the measures the decisions are taken from and not on ranking quality alone.

Serving-time averaging improves all three consumer measures at least as much for OC-SFT as for the off-the-shelf scorer (Figure 4). From $K = 1$ to $K = 10$, the pair-flip reduction is larger after OC-SFT, and the answer-flip reduction is equal. Under matched retention, OC-SFT also closes a larger share of the available overlap range. OC-SFT leads at every measured budget on all three consumers and on every collection from $K = 2$ onward, with a single tie on TREC-COVID at $K = 1$. OC-SFT reduces but does not eliminate presentation sensitivity. Averaging therefore further improves reproducibility after training, but produces a smaller gain in label accuracy (Table 14).

On reranking, OC-SFT at $K{=}\,10$ exceeds its $K{=}\,1$ result, while BSC at $K{=}\,10$ still trails OC-SFT at $K{=}\,1$. At $K{=}\,1$, OC-SFT reaches 0.832 retained-set overlap against 0.663 for single-order distillation and 0.737 for order-averaged distillation; at $K{=}\,10$ the three reach 0.950, 0.869 and 0.904; all three-seed means. OC-SFT therefore leads both baselines at matched K, although each baseline at $K{=}\,10$ exceeds OC-SFT at $K{=}\,1$. Figure 4 draws the same two variants on the queries they share, and at each K it sets OC-SFT's cutoff to the fraction the off-the-shelf scorer's F1-tuned cutoff retains, so both are thresholded to equal set size. We match them, as retained-set overlap rises with set size and the F1-tuned fraction itself moves with K. The same two OC-SFT points are 0.831 and 0.943 there. The tenfold comparison in Figure 4 holds against the untrained ensemble of Korikov et al. (2025) and does not extend to a trained one.

Pair flip counts how often the selected pair moves, not whether the movement crosses the gold label. We therefore classify each query by whether its top-scored response agrees with gold under all $M = 10$ random permutations, under none of them, or under some.

Table 14: **Gold crossing and label accuracy on response ranking.** *Gold crossing* is the share of queries whose top-scored response agrees with gold under some permutations and not others; *1 perm.* and *10 perm.* are label accuracy from a single permutation and from the ten-permutation average. Lower crossing and higher accuracy are better. The binary-gold mean excludes Nectar, whose gold is graded and whose off-the-shelf accuracy is near chance. Basis: $M = 10$ random permutations; off the shelf is a fixed evaluation, single-order distillation uses seed 42, and OC-SFT is the mean over seeds 42–44.

	Off the shelf			Single-order			OC-SFT		
Collection	Gold cross.↓	1 perm.	10 perm.	Gold cross.↓	1 perm.	10 perm.	Gold cross.↓	1 perm.	10 perm.
RewardBench-2	0.661	0.594	0.711	0.656	0.594	0.684	**0.295**	0.683	0.705
Nectar	0.714	0.225	0.295	0.698	0.261	0.300	**0.597**	0.276	0.318
PPE-MATH	0.654	0.754	0.834	0.523	0.806	0.846	**0.426**	0.812	0.843
PPE MMLU-Pro	0.727	0.662	0.691	0.770	0.663	0.668	**0.593**	0.667	0.677
RM-Bench	0.576	0.709	0.711	0.564	0.707	0.725	**0.496**	0.712	0.722
Binary-gold mean	0.654	0.680	0.737	0.628	0.693	0.731	**0.453**	0.719	0.737

Most pair flips occur on queries that cross the gold label (Table 14). Off the shelf, 0.654 of queries agree with gold under some permutations and not others, and 0.682 of all pair flips occur on those queries. The $K = 10$ average raises binary-gold mean label accuracy (Table 14).

The gain is concentrated in RewardBench-2 and PPE-MATH, while averaging adds almost nothing on RM-Bench or PPE MMLU-Pro. The mean therefore overstates how reliably averaging repairs a label. OC-SFT lowers crossing on all four binary-gold collections, and single-order distillation lowers it on three and raises it on PPE MMLU-Pro. The two trained variants are not matched on seeds, so we read the ordering rather than the size of the difference. At one presentation, OC-SFT approaches the accuracy of the ten-presentation off-the-shelf scorer, but further averaging improves label accuracy less for OC-SFT than for the off-the-shelf scorer. Averaging therefore adds little to label correctness once the scorer is trained, while on reproducibility it keeps composing (Figure 4). Nectar stays out of the mean, since its gold is the maximum of graded scores rather than a binary preference and the off-the-shelf scorer agrees under all $M = 10$ random permutations on 0.002 of its

queries. Its single-permutation accuracy of 0.225 is near what a random choice among its responses would reach.

D.7 Robustness to the first stage

Because all main reranking results use BM25 candidates, we test first-stage dependence by evaluating the same BM25-trained Qwen3-4B checkpoints on BGE-base-en-v1.5 (Xiao et al., 2024) and SPLADE++ ED (Formal et al., 2022) candidates without retraining.

OC-SFT retains the lowest τ-PSI while remaining near order-averaged distillation in quality under all three first stages (Table 15). On the 13-collection mean, OC-SFT ties order-averaged distillation under BM25, trails it by 0.004 under BGE, and leads it by 0.002 under SPLADE. OC-SFT also has the lowest per-collection τ-PSI on all 13 collections (Table 33).

Table 15: **Four variants under three first stages.** Each variant has separate nDCG@10 and τ-PSI columns, averaged over the same 13 collections. All trained variants use BM25 candidates during training, so the BGE and SPLADE rows test eval-only transfer. DL21–23 and Legal-A/B are excluded because no BGE or SPLADE candidate sets were available for them. Basis: Qwen3-4B, seed 42; BM25-trained checkpoints; per-collection results are in Table 33.

	Off the shelf		Single-order		Order-avg.		OC-SFT	
First stage	nDCG	τ-PSI	nDCG	τ-PSI	nDCG	τ-PSI	nDCG	τ-PSI
BM25	0.374	0.287	0.452	0.198	0.463	0.140	0.463	0.089
BGE (dense)	0.373	0.315	0.462	0.213	0.478	0.147	0.474	0.089
SPLADE (learned-sparse)	0.358	0.307	0.455	0.210	0.463	0.146	0.465	0.090

E Transfer to multi-document QA and response ranking

We repeat the measurements outside passage reranking, on multi-document question answering (App. E.1) and on response ranking (App. E.2), to test whether presentation dependence and the effect of training on it are specific to the reranking setting. Table 16 gives the serving-width effect for all three.

Table 16: **Serving-width effects differ by task and collection.** Rows compare OC-SFT at $B = 1$ and at the training width in parentheses. Width effect is training-width quality minus $B = 1$ quality. One-permutation effects include the order residual, whereas order-averaged effects average the training-width scores over $K = 10$ serving permutations. Basis: Qwen3-4B, seeds 42–44.

		Level		Width effect	
Task	Metric	$B = 1$	Training width	One permutation	Order-averaged
Reranking: 18 collections ($B = 20$)	nDCG@10	0.459	0.459	+0.000	+0.020
Multi-doc QA ($B = 10$), 3 collections	nDCG@10	0.918	0.961	+0.043	+0.047
Response ranking ($B = 4$), 5 collections					
RewardBench-2 (4 candidates)	nDCG@1	0.670	0.690	+0.020	+0.036
Nectar (7 candidates)	nDCG@1	0.610	0.617	+0.006	+0.026
PPE-MATH (8 candidates)	nDCG@1	0.711	0.807	+0.096	+0.132
PPE MMLU-Pro (8 candidates)	nDCG@1	0.602	0.675	+0.074	+0.076
RM-Bench (6 candidates)	nDCG@1	0.755	0.715	−0.040	−0.033

E.1 Multi-document question answering

Ten-order labels transfer part of the benefit of shared-context support selection, but not consistently across QA collections (Table 17). With the same frozen reader, a $B = 1$ student trained on those labels matches the $B = 10$ scorer on HotpotQA EM and recovers most of its F1 advantage, recovers only part of either gap on MuSiQue, and falls below the pointwise baseline on

Table 17: **Support ranking transfers zero-shot; reader accuracy depends on scorer context.** Support columns report the OC-SFT scorer at $B = 10$. Reader columns report the same frozen Granite-4.1-8B reader over the top five passages from OC-SFT, a pointwise scorer, or a $B = 1$ scorer trained on labels averaged over ten multi-context teacher orders. Higher nDCG@10, EM and F1 are better; lower τ-PSI is better. All metrics are averaged over the same permutations. Basis: Qwen3-4B, 200 questions per collection, seed 42, $M = 10$ random permutations.

	OC-SFT support		Reader: OC-SFT		Reader: pointwise		Reader: ten-order labels	
Collection	nDCG@10	τ-PSI	EM	F1	EM	F1	EM	F1
HotpotQA	0.964	0.082	0.621	0.743	0.575	0.684	0.623	0.734
2WikiMultiHopQA	0.975	0.102	0.431	0.546	0.427	0.532	0.398	0.524
MuSiQue	0.944	0.078	0.417	0.531	0.340	0.459	0.381	0.487

2WikiMultiHopQA. Teacher averaging can therefore distill some context benefit into a pointwise student, but it does not replace presenting the companion passages at serving time.

Table 18: **QA width effects are positive on all three collections and come from real companion passages.** Block (a) gives the mean one-permutation and order-averaged width effects by collection. Block (b) decomposes the one-permutation three-collection mean into informative-alternative, prompt-length/position and answer-skeleton terms. Width effect is $B = 10$ quality minus $B = 1$ quality. Basis: Qwen3-4B OC-SFT, three independently trained checkpoints; each served at both widths.

(a) Mean width effect by collection

Collection	One permutation	Order-averaged
HotpotQA	$+0.0345$	$+0.0377$
2WikiMultiHopQA	$+0.0441$	$+0.0496$
MuSiQue	$+0.0518$	$+0.0545$

(b) Three-collection one-permutation width decomposition

Three-collection component	Mean
Observed width effect	$+0.0435$
Informative alternatives	$+0.0482$
Prompt length and target position	$+0.0034$
Answer skeleton	-0.0081

Real companion passages, not the longer answer template, produce the QA width gain (Table 18). To separate companion content from the larger prompt and readout, we score the same target passage in four conditions: alone with one grade slot; with its nine real candidate passages and ten slots; with nine unrelated, length-matched filler passages and ten slots; and alone under the same ten-slot answer template. The total width effect compares real companions with the one-slot condition. Within that total, real companions minus fillers measures informative companion content, fillers minus the target-alone ten-slot condition measures prompt length and target position, and the two target-alone conditions isolate the answer template. Only the real-companion term is large and positive; the length/position term is small and the answer-template term is negative. A separate check replaces one $B = 1$ engine call with ten independently batched $B = 1$ calls; the three-collection mean changes only from $+0.0470$ to $+0.0459$. The gain therefore comes from information in the other candidate passages rather than from the larger template or request batching (§6.4).

E.2 Response ranking

The five collections contain four to eight candidates, with four requiring more than one $B = 4$ prompt. The pointwise control is trained at $B = 1$ on pointwise-teacher labels under the same Qwen3-4B evaluation protocol used for the batched student.

Shared-context serving raises mean nDCG@1 on both PPE collections, changes little on Nectar, and lowers it on RM-Bench; the PPE-MATH and RM-Bench signs transfer to Gemma-E4B (Table 19). Across the two PPE collections and RM-Bench, the ten-order-label and pointwise-label

$B = 1$ variants differ by at most 0.006 nDCG@1; shared-context serving then improves both PPE collections but degrades RM-Bench. The quality change therefore emerges when candidates share the serving prompt rather than from averaging shared-context information into the teacher labels. Gemma-E4B reproduces the positive PPE-MATH and negative RM-Bench effects, while the PPE MMLU-Pro interval covers zero.

Table 19: **Response-ranking width and design effects vary by collection and model family.** Block (a) reports mean fixed-weight width effects. Block (b) compares nDCG@1 levels for three training/serving designs under one Qwen3-4B evaluation-file readout; it omits Nectar, whose pointwise level is in Table 20. Block (c) repeats the batched-minus-pointwise comparison on Gemma-E4B with paired prompt-bootstrap intervals. Blocks (a) and (b) estimate different comparisons. Comparing prompts against UltraFeedback found no exact overlap for RewardBench-2 or either PPE collection. Basis: blocks (a–b) use Qwen3-4B and block (c) uses Gemma-E4B.

(a) Mean fixed-weight width effects

Collection	One-permutation effect	Order-averaged effect	Pool size
RewardBench-2	$+0.020$	$+0.036$	4
Nectar	$+0.006$	$+0.026$	7
PPE-MATH	$+0.096$	$+0.132$	8
PPE MMLU-Pro	$+0.074$	$+0.076$	8
RM-Bench	-0.040	-0.033	6

(b) Label construction and serving geometry, Qwen3-4B

Collection	Ten-order labels ($B = 1$)	Pointwise labels ($B = 1$)	OC-SFT ($B = 4$)
RewardBench-2	0.657	0.605	0.690
PPE-MATH	0.723	0.719	0.805
PPE MMLU-Pro	0.594	0.592	0.666
RM-Bench	0.748	0.754	0.721

(c) Family replication, Gemma-E4B

Collection	OC-SFT ($B = 4$) minus pointwise ($B = 1$)	95% CI
PPE-MATH	$+0.1094$	$[+0.0664, +0.1523]$
PPE MMLU-Pro	$+0.0098$	$[-0.0391, +0.0566]$
RM-Bench	-0.0648	$[-0.0897, -0.0407]$

Skywork-Reward-V2-Qwen3-4B (Liu et al., 2026) is a 4B pointwise reward model that scores each response independently; PairRM (Jiang et al., 2023) is a 0.4B pairwise reward model that compares two responses at a time (Table 20). Skywork has higher nDCG@1 on RewardBench-2, Nectar and RM-Bench, whereas OC-SFT leads on PPE-MATH and PPE MMLU-Pro. PairRM falls below the pointwise control on all five collections; the paired interval excludes zero on four. PairRM was trained on six large preference datasets, including UltraFeedback, and Skywork also uses much heavier preference supervision than our roughly 30K prompts. Architecture, scale and training data all differ at once, so the comparison cannot attribute the gap to any one of them.

Table 20: **PairRM and Skywork against matched Qwen3-4B controls.** Each row gives matched nDCG@1 levels; PairRM contrasts use paired query bootstraps with 20,000 resamples. Basis: seed 42; matched query IDs per collection; UltraFeedback is excluded because it is train-on-test for PairRM.

		nDCG@1				PairRM − pointwise		OC-SFT − PairRM	
Collection	n	Pointwise	PairRM	Skywork	OC-SFT	Estimate	95% CI	Estimate	95% CI
Nectar	434	0.5975	0.5914	0.6467	0.6152	-0.0061	$[-0.0430, +0.0315]$	$+0.0238$	$[-0.0115, +0.0591]$
RewardBench-2	1,763	0.6052	0.4974	0.8174	0.6903	-0.1078	$[-0.1361, -0.0788]$	$+0.1929$	$[+0.1651, +0.2201]$
PPE-MATH	512	0.7188	0.4707	0.7656	0.8047	-0.2480	$[-0.3047, -0.1914]$	$+0.3340$	$[+0.2793, +0.3867]$
PPE MMLU-Pro	512	0.5918	0.4961	0.6445	0.6660	-0.0957	$[-0.1523, -0.0371]$	$+0.1699$	$[+0.1133, +0.2246]$
RM-Bench	1,327	0.7543	0.6353	0.8237	0.7212	-0.1191	$[-0.1492, -0.0889]$	$+0.0859$	$[+0.0543, +0.1176]$

F CONSUMER OUTPUTS AND REPORTED SCORES

We apply a frozen threshold (App. F.1), a frozen reader (App. F.3) and a fixed chosen/rejected-pair selector (App. F.4) to the same $M = 10$ random permutations, relate the retained set to standard ranking metrics (App. F.2), and recompute a reported benchmark score (App. F.5), to test whether presentation dependence propagates from rankings to downstream decisions and reported quality. Holding each consumer fixed isolates variation caused by the scorer because the consumer itself does not change.

F.1 A FROZEN THRESHOLD: THE RETAINED SET

For each pair of scorer permutations, let A and B be the documents admitted by the same frozen threshold. *Retained-set overlap* averages the Jaccard index $|A \cap B|/|A \cup B|$ over all permutation pairs; it equals one when every permutation retains the same set.

The retained-set gain transfers across model families (Table 21). Every dense base improves on the 18-collection mean, with positive collection-level gains in 196 of 198 cells; the mixture-of-experts base improves on all 18 collections. What the threshold retains is therefore made more reproducible by the objective rather than by the primary base.

Table 21: **Retained-set overlap gain of OC-SFT over single-order distillation by base model.** Each gain is an equal-weight collection mean. Single-order distillation is the baseline because the off-the-shelf fixed threshold degenerates to retain-all on TREC-COVID. The 95% CI column gives paired collection-bootstrap intervals; the last column counts positive collection-level gains. Basis: 18 reranking collections at a frozen per-collection F1 threshold, seed 42, independently selected checkpoints on each base.

Base	Gain	95% CI	Improving
Qwen3-1.7B	+0.369	[+0.298, +0.423]	17/18
Qwen3-4B	+0.169	[+0.152, +0.186]	18/18
Qwen3-8B	+0.254	[+0.233, +0.274]	18/18
Qwen3-14B	+0.248	[+0.210, +0.282]	18/18
Qwen3-32B	+0.264	[+0.237, +0.292]	18/18
Gemma-E2B	+0.186	[+0.173, +0.201]	18/18
Gemma-E4B	+0.116	[+0.100, +0.131]	18/18
Gemma-31B	+0.135	[+0.088, +0.182]	17/18
Granite-3B	+0.174	[+0.149, +0.202]	18/18
Granite-8B	+0.137	[+0.123, +0.152]	18/18
Granite-30B	+0.216	[+0.197, +0.235]	18/18
Gemma-26B-A4B MoE	+0.106	[+0.085, +0.135]	18/18

F.2 RANKING METRICS AND THE RETAINED SET

Retained-set overlap measures a real loss, and no ranking metric substitutes for it (Table 22). For both single-order distillation and OC-SFT, per-query retained-set F1 varies about as much as nDCG@10 across presentations (Figure 3). The loss is concentrated near the cutoff, where churned documents have relevance density 0.215, versus 0.438 among documents retained under both permutations. The ordering τ-PSI induces agrees with the ordering overlap induces more often than nDCG@10's does, but the agreement is incomplete. Collection averaging reduces the apparent spread, so neither a ranking metric nor a collection mean shows how a threshold's output changes under reshuffling.

F.3 A FROZEN READER: THE VERDICT AND THE ANSWER

At 4B, OC-SFT lowers verdict and answer flip while reader accuracy remains level (Table 23). The default reader is a frozen Granite-4.1-8B decoding greedily into three constrained claim-verification labels. A frozen Qwen3-4B reader repeats the comparison on both claim-verification and

Table 22: **Retained-set overlap by collection for seven single-permutation systems within 0.022 nDCG@10.** The quality and overlap winners differ on 11 of 18 collections. Excluding GPT-5.4, whose τ-PSI is not reported, median pairwise disagreement with overlap is 0.067 for τ-PSI and 0.267 for nDCG@10. The six-system discordance calculation uses seed-42 nDCG@10 and τ-PSI, with jina at matched $B = 20$. *F1 spread* is a query's mean retained-set F1 minus its worst permutation, averaged over queries, so the set a threshold admits is about as permutation-sensitive as the ranking it is read from. Basis: three-seed means for trained rows and single evaluations for external rows, at the frozen per-collection F1 thresholds of Table 1; per-collection quality and instability are in Tables 32, 35 and 36.

Collection	*GPT-5.4*	OC-SFT	Order-avg.	Perm. aug.	DebiasFirst	Single-order	*jina-v3*
DL19	0.779	0.882	0.827	0.832	0.835	0.732	0.769
DL20	0.782	0.881	0.812	0.827	0.826	0.743	0.721
DL21	0.846	0.899	0.814	0.832	0.832	0.692	0.756
DL22	0.685	0.825	0.785	0.774	0.774	0.638	0.685
DL23	0.617	0.814	0.701	0.747	0.741	0.634	0.621
Touché-2020	0.832	0.832	0.757	0.795	0.773	0.622	0.670
FiQA	0.640	0.850	0.723	0.750	0.748	0.661	0.662
NFCorpus	0.738	0.862	0.770	0.796	0.791	0.679	0.666
ArguAna	0.607	0.615	0.539	0.544	0.540	0.484	0.618
Climate-FEVER	0.624	0.798	0.707	0.704	0.711	0.621	0.589
TREC-COVID	0.831	0.933	0.887	0.901	0.902	0.778	0.791
DBPedia	0.696	0.836	0.730	0.756	0.753	0.653	0.631
SciFact	0.806	0.886	0.785	0.787	0.796	0.744	0.688
Signal-1M	0.597	0.813	0.731	0.721	0.726	0.634	0.625
TREC-NEWS	0.723	0.841	0.728	0.741	0.760	0.644	0.655
Robust04	0.739	0.859	0.770	0.784	0.782	0.694	0.665
Legal-A	0.579	0.779	0.604	0.671	0.651	0.544	0.581
Legal-B	0.606	0.828	0.701	0.720	0.715	0.610	0.605
Mean overlap ↑	0.707	**0.835**	0.743	0.760	0.759	0.656	0.667
Summary measures, 18-collection means							
nDCG@10 ↑	0.468	0.459	0.455	0.455	0.454	0.449	0.447
Retained-set F1 ↑	0.415	0.407	0.410	0.406	0.409	0.405	0.420
Retained-set F1 spread ↓	0.081	**0.040**	0.057	0.056	0.055	0.071	0.087
τ-PSI ↓	—	**0.083**	0.130	0.129	0.128	0.209	0.177

all three question-answering collections. *Verdict flip* and *answer flip* are mean pairwise disagreement rates over the $M = 10$ random permutations. The reductions are positive under both readers on claim verification and QA, while all reported reader-accuracy intervals cover zero. The same-evidence controls are positive for claim verification and QA, and their magnitudes fall within the range of the all-query effects. The lower flip rates therefore do not require a change in which evidence reaches the reader. EMA closes more of the rank-level gap than augmentation but loses 0.023 QA exact match; together, these controls show that OC-SFT's rank-level gain reaches the reader output without a measurable accuracy trade-off (Tables 10 and 23).

Table 23: **Verdict- and answer-flip reductions and controls at 4B.** Positive estimates favor the variant after the arrow: lower flip for flip rows and higher accuracy for accuracy rows. Block (a) reports SD where repeated fits are available; block (b) reports query-paired 95% intervals at one fixed fit. In block (c), percentages are the share of queries on which both variants select the same top five. Block (d) uses a query-paired interval for Climate-FEVER and hierarchical query-within-collection intervals for the QA aggregates. Basis: Granite-4.1-8B reader unless Qwen3-4B is named; QA rows use 200 questions per collection.

(a) Repeated-fit flip and accuracy estimates

Contrast	Collection / output	Estimate	SD
Single-order → OC-SFT	Climate-FEVER	+0.0263	0.0031
Single-order → OC-SFT	SciFact	+0.0251	0.0014
Single-order → OC-SFT, second reader	Climate-FEVER	+0.0304	0.0025
Perm. aug. → EMA	QA exact match	−0.0228	0.0029

(b) Flip reductions with query-paired intervals

Contrast	Collection / output	n	Estimate	95% CI
Single-order → OC-SFT, second reader	SciFact	300	+0.0368	[+0.0207, +0.0532]
Single-order → OC-SFT	HotpotQA answer	200	+0.0487	[+0.0283, +0.0698]
Single-order → OC-SFT	2Wiki answer	200	+0.0459	[+0.0221, +0.0707]
Single-order → OC-SFT	MuSiQue answer	200	+0.0700	[+0.0427, +0.0979]
Single-order → OC-SFT, second reader	HotpotQA answer	200	+0.0274	[+0.0054, +0.0494]
Single-order → OC-SFT, second reader	2Wiki answer	200	+0.0891	[+0.0643, +0.1143]
Single-order → OC-SFT, second reader	MuSiQue answer	200	+0.0823	[+0.0538, +0.1122]

(c) Same-evidence point controls

Contrast	Collection / output	Matched-query share	Estimate
Single-order → OC-SFT, same top 5	Climate-FEVER verdict	17.5%	+0.030
Single-order → OC-SFT, same top 5	QA answer	45.2%	+0.061

(d) Reader accuracy

Contrast	Collection / output	n	Estimate	95% CI
Single-order → OC-SFT	Climate-FEVER verdict	1,381	−0.0036	[−0.0111, +0.0038]
Single-order → OC-SFT	QA exact match	200×3	−0.0058	[−0.0153, +0.0035]
Single-order → OC-SFT	QA F1	200×3	−0.0016	[−0.0104, +0.0071]

F.4 A PREFERENCE MODEL: THE SELECTED PAIR

A preference pipeline consumes both ends of a ranking, so *pair flip* is the mean pairwise rate at which permutations select different chosen and rejected responses.

OC-SFT is the most reproducible trained variant on every response-ranking collection (Table 24). On each collection, it changes the chosen/rejected pair less often than every other trained row, so the rank-level stability gain reaches the preference decision rather than only the complete ranking. Gold crossing and serving-time K sweeps for this consumer are reported in App. D.6.

Table 24: **Preference-pair flip on response ranking.** Lower is better. Basis: Qwen3-4B seed 42; Nectar uses the deduplicated 434-prompt subset.

Collection	Off the shelf	Single-order	Order-avg.	Perm. aug.	DebiasFirst	OC-SFT
RewardBench-2	0.763	0.757	0.496	0.447	0.441	**0.356**
Nectar	0.905	0.872	0.718	0.704	0.700	**0.650**
PPE-MATH	0.920	0.914	0.796	0.794	0.792	**0.744**
PPE MMLU-Pro	0.930	0.940	0.845	0.835	0.837	**0.818**
RM-Bench	0.869	0.861	0.769	0.776	0.776	**0.739**

F.5 A REPORTED BENCHMARK SCORE

A reported benchmark score may use one permutation, whereas τ-PSI and retained-set overlap aggregate all ten. We therefore test whether a single permutation supports the same cross-system quality conclusion as the order-marginal score. Skywork (Liu et al., 2026) is pointwise and has one

fixed nDCG@1 per collection, whereas OC-SFT has one score per permutation; the system with the higher collection-level nDCG@1 is the leader.

Permutation range exceeds the mean OC-SFT–Skywork gap on two collections without reversing the observed leader (Table 25). The range is wider than the absolute gap on Nectar and PPE MMLU-Pro, but no one of the $M = 10$ random permutations reverses the leader. Permutation choice therefore changes the apparent margin rather than the observed ordering on these runs. Only PPE MMLU-Pro's paired query-bootstrap interval covers zero, so query sampling leaves that comparison unresolved as well. We report the order-marginal OC-SFT–Skywork gap rather than selecting one arbitrary permutation; the paired bootstrap quantifies the remaining query-sampling uncertainty.

Table 25: **Permutation range exceeds the mean system gap on two collections.** Each OC-SFT permutation is aggregated over the prompts scored by pointwise Skywork, so both sides share a basis. Range is max minus min of OC-SFT's nDCG@1 across M=10 permutations. Bold marks ranges larger than the absolute system gap. Perm. SD is the standard deviation across permutations. Gap is OC-SFT minus Skywork; its intervals are paired prompt bootstraps and carry query sampling rather than permutation range. Basis: Qwen3-4B seed 42; Nectar uses the deduplicated 434-prompt subset.

Collection	n	Perm. SD	Range	Gap	95% CI
RewardBench-2	1,763	0.0065	0.0199	−0.136	[−0.154, −0.118]
Nectar	434	0.0162	**0.0476**	−0.031	[−0.060, −0.002]
PPE-MATH	512	0.0102	0.0332	+0.047	[+0.014, +0.080]
PPE MMLU-Pro	512	0.0129	**0.0410**	+0.019	[−0.017, +0.055]
RM-Bench	1,327	0.0109	0.0317	−0.110	[−0.128, −0.092]

G SERVING WIDTH AND SYSTEM TRADE-OFFS

We measure how serving width affects quality and stability (App. G.1), decompose the width effect into its sources (App. G.2), test tolerance to a mismatch between training and serving width (App. G.3), and report the range that is safe to serve (App. G.4) and the latency of batched serving (App. G.5).

G.1 PASSAGE-RERANKING WIDTH AND ESTIMATOR CONTROLS

A single permutation does not measurably improve with serving width on passage reranking, while order averaging adds 0.020 nDCG@10 (Table 26). The order-averaged comparison averages ten $B = 20$ permutations and compares them with one $B = 1$ pass, whereas the one-permutation column compares one pass at $B = 20$ with one at $B = 1$. Matching the $B = 1$ estimator with ten independently rebatched passes leaves the order-averaged result nearly unchanged (Table 26). Each $B = 1$ prompt still contains one candidate, so this intervention changes engine-call batching, meaning how many prompts the inference engine groups into one call, rather than what shares a prompt. The $+0.020$ effect is therefore not an artifact of averaging only the $B = 20$ scores.

Engine-call batching has a much smaller effect than candidate order (Table 26). The two batching-only readings are close despite different serving widths, while changing candidate order at $B = 20$ produces lower retained-set overlap; these losses are directional controls rather than additive components because changing order also changes engine-call composition.

G.2 DECOMPOSITION OF THE WIDTH EFFECT

Outside ArguAna, the order-averaged width effect is almost entirely companion content (Table 27). ArguAna instead derives most of its effect from length, slot and skeleton terms.

G.3 TRAINING WIDTH AND TRAIN-SERVE MISMATCH

Training at $B = 20$ makes the weights width-tolerant, whereas a $B = 1$-trained student loses quality when served at $B = 20$ (Table 28). Only the off-the-shelf scorer and the $B = 1$-trained student have width-effect intervals excluding zero. In the train-serve comparison, the $B = 20$-trained student transfers to $B = 1$, while the $B = 1$-trained student loses 0.050 nDCG@10 on a wide pass

Table 26: **The $+0.020$ passage-reranking gain survives estimator matching, while engine-call batching has little effect on retained-set overlap.** Block (a) reports the one-permutation and order-averaged width effects, including a matched $B = 1$ estimator built from ten independently rebatched passes. Block (b) holds candidate order fixed and varies engine-call batching, then contrasts it with the candidate-order perturbation. Basis: block (a) uses the 18 reranking collections at seed 42; block (b) uses the 16 public ones at seed 42 and the frozen-threshold protocol of App. F.

(a) nDCG@10 width effect

Comparison	Collection basis	Effect	95% CI
One permutation	18 reranking	$+0.0001$	$[-0.0133, +0.0177]$
Order-averaged, original	18 reranking	$+0.0201$	$[+0.0063, +0.0429]$
Order-averaged, matched $B = 1$	18 reranking	$+0.0191$	$[+0.0050, +0.0424]$
Order-averaged, original	Excluding ArguAna	$+0.0098$	$[+0.0051, +0.0147]$
Order-averaged, matched $B = 1$	Excluding ArguAna	$+0.0088$	$[+0.0035, +0.0144]$

(b) Retained-set overlap under matched perturbations

Serving width	Perturbation	Overlap
$B = 1$	Engine-call batching	0.9797
$B = 20$	Engine-call batching, fixed candidate order	0.9848
$B = 20$	Candidate order	0.8353

Table 27: **Decomposition of the width effect.** At fixed Qwen3-4B OC-SFT weights, q_{alone} scores a target alone under one answer slot, q_{comp} adds its 19 real companions, q_{fill} substitutes 19 unrelated length-matched fillers, and q_{skel} returns the target alone under the full 20-slot template. Width is $q_{\text{comp}} - q_{\text{alone}}$, content is $q_{\text{comp}} - q_{\text{fill}}$, length plus slot is $q_{\text{fill}} - q_{\text{skel}}$, and skeleton is $q_{\text{skel}} - q_{\text{alone}}$. The upper block uses the first-stage permutation; the lower block averages the $B = 20$ scores over $K = 10$ serving permutations. Positive entries mean the wider serving width ranks better. Content, length plus slot, and skeleton sum to width up to rounding. Starred cells have query-paired 95% bootstrap intervals clear of zero. Basis: Qwen3-4B OC-SFT, seed 42, 18 reranking collections.

Collection basis	Width	Content	Length + slot	Skeleton
First-stage order				
18 reranking	-0.0002	-0.0072	$+0.0039$	$+0.0031$
Excluding ArguAna	-0.0069	-0.0047	-0.0016	-0.0007
ArguAna	$+0.1144^{*}$	-0.0496^{*}	$+0.0965^{*}$	$+0.0675^{*}$
Averaged over $K = 10$ serving permutations				
18 reranking	$+0.0194$	$+0.0085$	$+0.0026$	$+0.0083$
Excluding ArguAna	$+0.0089$	$+0.0086$	$+0.0007$	-0.0004
ArguAna	$+0.1979^{*}$	$+0.0071^{*}$	$+0.0341^{*}$	$+0.1567^{*}$

and recovers it only by averaging ten evaluations. The same train-serve question for the teacher's labelling width is reported in Table 11.

Training attenuates the score-level width effect without removing it, and the consistency penalty adds no distinguishable further reduction (Table 29). Single-order distillation and permutation augmentation reduce $|\mu_B - \mu_1|$, but augmentation and OC-SFT differ by only 0.013 score standard deviations with an interval covering zero. Although training attenuates its magnitude, OC-SFT remains well above zero and therefore has not become pointwise.

G.4 Serving-width operating range

Quality remains stable from $B = 1$ through $B = 30$ (Table 30). Across that range, mean nDCG@10 over the 18 reranking collections changes little, and τ-PSI remains near its training-width value. At $B = 34$ both measures degrade, and at $B = 40$ and $B = 50$ they deteriorate further.

Table 28: **Training width determines tolerance to serving-width changes.** Block (a) measures τ-PSI at $B{=}\ 20$ and the single-pass nDCG@10 change from serving the same seed-42 weights at $B{=}\ 1$ and $B{=}\ 20$; intervals are paired 95% bootstraps. Block (b) crosses training and serving width using three-seed means; its final column averages ten $B = 20$ permutations. $B{=}\ 1$ τ-PSI entries are zero by construction. Basis: Qwen3-4B, 18 reranking collections, equal-weight collection mean.

(a) Fixed-weight width sensitivity, seed 42

Variant	τ-PSI	Width effect	95% CI
Not trained at the serving width			
Off the shelf	0.298	-0.062	$[-0.082, -0.043]$
Student trained at $B = 1$	0.222	-0.048	$[-0.062, -0.036]$
Trained at $B = 20$			
Order-avg.	0.139	-0.004	$[-0.012, +0.004]$
DebiasFirst	0.129	-0.005	$[-0.015, +0.005]$
Perm. aug.	0.122	-0.006	$[-0.016, +0.003]$
OC-SFT	**0.085**	$+0.000$	$[-0.013, +0.018]$

(b) Train-serve width comparison, three-seed means

	Serve $B = 1$		Serve $B = 20$, one pass		Serve $B = 20$, $K = 10$
Train	nDCG@10	τ-PSI	nDCG@10	τ-PSI	nDCG@10
$B = 20$ (OC-SFT)	0.459	0	0.459	0.083	0.479
$B = 1$-trained student	0.466	0	0.415	0.227	0.470

Table 29: **Magnitude of the score-level width effect across training variants.** Entries are collection-first mean $|\mu_B - \mu_1|$ normalized by pooled across-document score SD; they include companion content and prompt geometry. The final row isolates the consistency penalty. Order-averaged distillation is omitted because it changes the teacher targets. Basis: 18 reranking collections, Qwen3-4B, seed 42; $B = 20$ averages $K = 10$ serving permutations, $B = 1$ scores each document alone, and intervals bootstrap collections.

Variant / contrast	Level or difference	95% CI
Off the shelf	0.521	[0.449, 0.593]
Single-order	0.300	[0.246, 0.383]
Perm. aug.	0.245	[0.198, 0.323]
OC-SFT	0.232	[0.175, 0.330]
Augmentation minus OC-SFT	$+0.013$	$[-0.012, +0.034]$

Table 30: **OC-SFT quality is stable through $B = 1$ to $B = 30$.** The selected $B = 20$-trained OC-SFT checkpoint is served at nine widths; higher nDCG@10 and lower τ-PSI are better. The training width is bold in the header. Basis: Qwen3-4B, seed 42, $M = 10$ random permutations for $B > 1$; at $B = 1$, τ-PSI is zero by construction.

Serve B	1	5	10	**20**	25	30	34	40	50
18-collection nDCG@10 ↑	0.459	0.453	0.454	**0.459**	0.461	0.460	0.446	0.345	0.263
18-collection τ-PSI ↓	0	0.094	0.091	**0.085**	0.083	0.094	0.135	0.233	0.275

G.5 Deployment and system trade-offs

Shared-context students match RankZephyr quality at lower latency while retaining per-document scores (Figure 8). All model-family points in the blue cluster are OC-SFT students, not off-the-shelf checkpoints. The Qwen3-Reranker-4B point starts from a purpose-built reranking checkpoint; our OC-SFT adaptation raises its public-16 nDCG@10 by 0.05 and lowers τ-PSI by 72% (Table 34). The 30B-class students similarly match RankZephyr's mean quality at lower latency while emitting per-document scores. jina and mxbai are external, purpose-built rerankers trained on substantially broader corpora and with specialized procedures, including cross-system hard-negative mining for jina and reinforcement, contrastive and preference learning for mxbai (Wang et al., 2025a; Li et al., 2026). They are therefore reference systems rather than supervision-matched baselines: mxbai leads on native-geometry quality and jina on latency, while RankZephyr remains more order-sensitive.

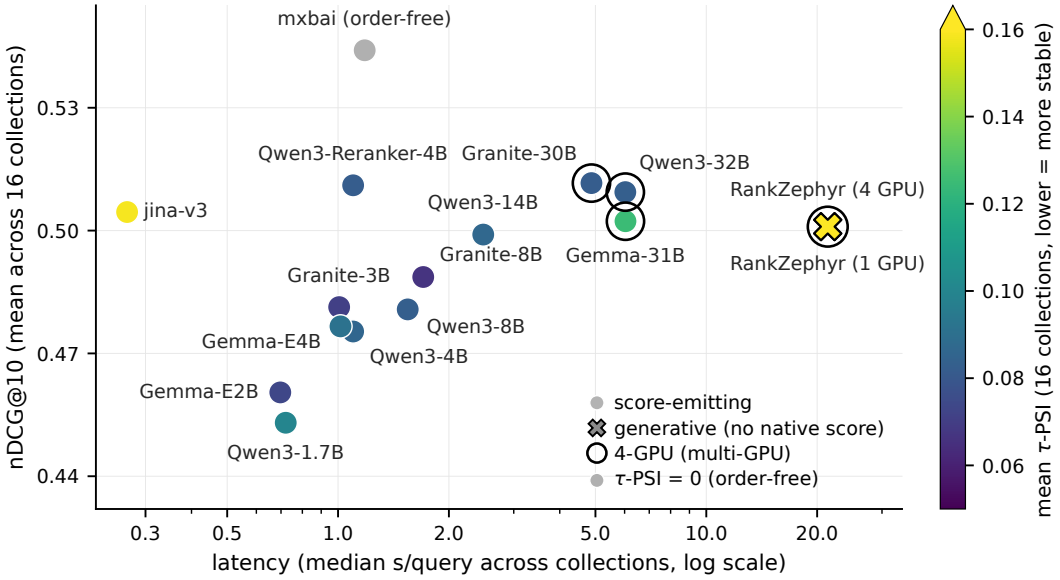


Figure 8: **The system frontier across 16 public reranking collections.** The horizontal axis is median per-query latency, the vertical axis mean nDCG@10, color mean τ-PSI, and marker shape whether the system emits a per-document score. The model-family points are trained OC-SFT checkpoints on the named bases. jina, mxbai and RankZephyr are external rerankers shown at their stated serving configurations. Gray denotes τ-PSI zero by construction for the order-free point. nDCG@10 and τ-PSI are equal-weight means and latency is the median of per-collection medians over 16 public reranking collections. Ringed points use four L40S GPUs rather than one.

Candidate batching roughly halves per-query latency at concurrency 1 (Table 31). Across the 16 public collections, the median reduction relative to $B = 1$ is 51.3% at $B = 10$ and 45.4% at $B = 20$. We retain $B = 20$ to match the training geometry, while $B = 10$ offers lower median latency except on Signal-1M and DL21–23. The batching reduction compounds with serving one permutation rather than ten (Figure 4).

Table 31: **Candidate batching roughly halves per-query latency.** Collection rows report seconds per query and the reduction relative to $B = 1$. Summary rows are equal-collection medians; brackets give 95% bootstrap intervals for the reduction. Basis: Qwen3-4B OC-SFT, seed 42, 18 collections, one L40S, warm vLLM and client concurrency 1. Each collection is timed three times on the same GPU, with the width order varied across repetitions. These within-model timings use query batch size 1 and are separate from the system-level latency comparison in Figure 8.

Collection	$B = 1$	$B = 10$	$B = 20$	$B = 10 \downarrow$	$B = 20 \downarrow$
DL19	2.223	1.020	1.086	54.1%	51.2%
DL20	2.237	1.032	1.124	53.9%	49.7%
TREC-COVID	2.214	1.111	1.255	49.8%	43.3%
NFCorpus	2.249	1.092	1.226	51.5%	45.5%
Touché-2020	2.404	1.176	1.315	51.1%	45.3%
DBPedia-Entity	2.235	1.036	1.100	53.6%	50.8%
SciFact	2.272	1.145	1.304	49.6%	42.6%
Signal-1M	2.098	0.590	0.557	71.9%	73.4%
TREC-News	2.241	1.181	1.354	47.3%	39.6%
Robust04	2.285	1.230	1.418	46.2%	37.9%
FiQA	2.296	1.193	1.373	48.1%	40.2%
ArguAna	2.345	1.228	1.372	47.6%	41.5%
Climate-FEVER	2.273	1.144	1.317	49.7%	42.0%
DL21	2.188	0.970	0.913	55.7%	58.3%
DL22	2.206	0.974	0.914	55.8%	58.6%
DL23	2.172	0.944	0.914	56.5%	57.9%
Legal-A	2.278	1.134	1.274	50.2%	44.1%
Legal-B	2.264	1.111	1.243	50.9%	45.1%
Public-16 median	2.239	1.101	1.240	51.3% [48.9, 54.1]	45.4% [42.0, 51.7]
Primary-18 median	2.245	1.111	1.249	51.0% [49.7, 54.0]	45.2% [42.3, 51.0]

H PER-COLLECTION AND CROSS-BASE REFERENCE RESULTS

Finally, we report per-collection results for the Qwen3-4B baselines (App. H.1), under alternative first stages (App. H.2), against specialized and external rerankers (App. H.3), and across twelve base models (App. H.4), so that each aggregate finding can be checked collection by collection and its exceptions identified.

H.1 QWEN3-4B PER-COLLECTION REFERENCE

Table 32: **Full per-collection reranking results for the Qwen3-4B student and its baselines.** Higher nDCG@10 and lower τ-PSI are better. BM25 has zero τ-PSI by construction; GPT-5.4 has no reported τ-PSI because tied grades leave its relative ordering to the tie-break, which is document identifier for nDCG@10, as noted in Table 1. Basis: seed 42 and $M = 10$ random permutations; *Public 16* excludes Legal-A and Legal-B, whose candidate lists are precomputed. Body Table 1 reports the mean over all 18 collections.

Method	Public 16	DL19	DL20	DL21	DL22	DL23	Touché	FiQA	NFCorpus	ArguAna	Climate	T-COVID	DBPedia	SciFact	Signal1M	T-NEWS	Robust04	Legal-A	Legal-B
nDCG@10, higher is better																			
First-stage retrieval without reranking																			
BM25 retrieval	0.39	0.51	0.49	0.45	0.27	0.26	0.44	0.24	0.34	0.40	0.17	0.59	0.32	0.68	0.33	0.40	0.41	0.30	0.32
Prompted closed model ($B = 20$)																			
GPT-5.4	0.49	0.68	0.65	0.65	0.45	0.42	0.23	0.37	0.38	0.40	0.19	0.85	0.39	0.75	0.27	0.50	0.57	0.36	0.31
Qwen3-4B																			
Off the shelf	0.38	0.60	0.55	0.54	0.39	0.33	0.22	0.25	0.31	0.21	0.16	0.73	0.33	0.51	0.22	0.33	0.45	0.29	0.24
CapCal	0.38	0.62	0.55	0.55	0.39	0.34	0.22	0.25	0.31	0.20	0.16	0.72	0.34	0.50	0.22	0.34	0.45	0.29	0.24
Single-order	0.46	0.72	0.67	0.67	0.45	0.43	0.26	0.35	0.35	0.26	0.19	0.75	0.40	0.70	0.29	0.42	0.52	0.34	0.29
Order-avg.	0.47	0.73	0.68	0.67	0.45	0.44	0.29	0.35	0.36	0.28	0.21	0.75	0.41	0.71	0.29	0.44	0.52	0.36	0.31
Perm. aug.	0.47	0.73	0.67	0.67	0.46	0.45	0.27	0.35	0.36	0.26	0.20	0.75	0.41	0.71	0.28	0.42	0.52	0.35	0.31
DebiasFirst	0.47	0.72	0.68	0.67	0.46	0.44	0.27	0.35	0.36	0.24	0.20	0.75	0.41	0.71	0.29	0.43	0.52	0.35	0.31
OC-SFT	0.48	0.73	0.68	0.68	0.47	0.45	0.28	0.37	0.36	0.31	0.19	0.74	0.42	0.71	0.30	0.42	0.52	0.35	0.31
τ-PSI, lower is better																			
Qwen3-4B																			
Off the shelf	0.29	0.30	0.30	0.31	0.32	0.33	0.29	0.26	0.28	0.27	0.27	0.28	0.32	0.30	0.33	0.27	0.26	0.34	0.34
CapCal	0.29	0.30	0.30	0.31	0.32	0.33	0.29	0.26	0.28	0.27	0.27	0.28	0.32	0.30	0.33	0.27	0.27	0.34	0.34
Single-order	0.20	0.19	0.18	0.18	0.21	0.21	0.18	0.18	0.21	0.22	0.19	0.18	0.22	0.23	0.23	0.17	0.19	0.23	0.23
Order-avg.	0.14	0.12	0.12	0.12	0.14	0.14	0.11	0.12	0.15	0.17	0.14	0.11	0.17	0.18	0.17	0.12	0.14	0.14	0.14
Perm. aug.	0.12	0.10	0.11	0.10	0.12	0.12	0.10	0.10	0.14	0.16	0.13	0.09	0.14	0.16	0.15	0.11	0.14	0.12	0.12
DebiasFirst	0.13	0.11	0.11	0.10	0.12	0.12	0.10	0.11	0.15	0.17	0.14	0.09	0.14	0.18	0.16	0.12	0.15	0.12	0.12
OC-SFT	**0.09**	**0.07**	**0.07**	**0.06**	**0.08**	**0.08**	**0.07**	**0.07**	**0.09**	**0.13**	**0.09**	**0.06**	**0.10**	**0.10**	**0.11**	**0.08**	**0.10**	**0.08**	**0.08**

H.2 Alternative first stages

Table 33: **Per-collection results under dense and learned-sparse first stages.** Each variant has separate nDCG@10 and τ-PSI columns. The three trained variants use BM25 candidates during training; the BGE and SPLADE results therefore measure transfer to new first stages without retraining. Higher quality and lower instability are better. Basis: Qwen3-4B, seed 42 and first-stage-specific judged sets; the **Mean (13)** row reproduces the BGE and SPLADE rows in Table 15, whose first row reports BM25.

	BGE (dense)								SPLADE++ ED (learned-sparse)							
	Off the shelf		Single-order		Order-avg.		OC-SFT		Off the shelf		Single-order		Order-avg.		OC-SFT	
Collection	nDCG	τ-PSI	nDCG	τ-PSI	nDCG	τ-PSI	nDCG	τ-PSI	nDCG	τ-PSI	nDCG	τ-PSI	nDCG	τ-PSI	nDCG	τ-PSI
DL19	0.573	0.336	0.740	0.202	0.758	0.131	0.740	0.073	0.586	0.333	0.739	0.209	0.727	0.134	0.741	0.075
DL20	0.533	0.328	0.720	0.193	0.729	0.130	0.723	0.074	0.513	0.331	0.719	0.197	0.718	0.129	0.713	0.074
Touché-2020	0.219	0.300	0.262	0.187	0.267	0.113	0.273	0.070	0.193	0.287	0.247	0.178	0.241	0.114	0.246	0.079
FiQA	0.256	0.304	0.360	0.210	0.377	0.131	0.382	0.070	0.250	0.292	0.354	0.202	0.356	0.128	0.365	0.070
NFCorpus	0.281	0.324	0.355	0.245	0.369	0.190	0.371	0.118	0.266	0.317	0.331	0.235	0.347	0.186	0.344	0.117
ArguAna	0.211	0.308	0.250	0.239	0.292	0.175	0.316	0.127	0.205	0.295	0.239	0.230	0.280	0.172	0.310	0.127
Climate-FEVER	0.160	0.284	0.179	0.200	0.211	0.143	0.188	0.079	0.152	0.280	0.181	0.201	0.205	0.147	0.184	0.086
DBPedia	0.370	0.314	0.459	0.196	0.473	0.147	0.463	0.084	0.360	0.321	0.443	0.213	0.460	0.153	0.453	0.088
SciFact	0.529	0.319	0.703	0.244	0.713	0.180	0.718	0.103	0.511	0.311	0.693	0.235	0.706	0.176	0.711	0.104
Signal-1M	0.192	0.346	0.250	0.246	0.267	0.174	0.262	0.115	0.197	0.340	0.277	0.241	0.272	0.172	0.274	0.113
TREC-COVID	0.730	0.323	0.758	0.210	0.763	0.116	0.757	0.070	0.668	0.305	0.729	0.200	0.719	0.112	0.724	0.068
TREC-NEWS	0.335	0.310	0.412	0.199	0.426	0.137	0.420	0.080	0.298	0.296	0.400	0.194	0.412	0.138	0.407	0.084
Robust04	0.462	0.293	0.557	0.201	0.567	0.142	0.555	0.091	0.460	0.285	0.563	0.193	0.571	0.139	0.568	0.089
Mean (13)	**0.373**	**0.315**	**0.462**	**0.213**	**0.478**	**0.147**	**0.474**	**0.089**	**0.358**	**0.307**	**0.455**	**0.210**	**0.463**	**0.146**	**0.465**	**0.090**

H.3 SPECIALIZED AND EXTERNAL RERANKERS

On Qwen3-Reranker-4B, OC-SFT matches order-averaged distillation in quality and lowers mean instability (Table 34). Serving `jina-reranker-v3` at matched $B = 20$ lowers its quality and raises its τ-PSI relative to its native single-context configuration.

Table 34: **Purpose-built rerankers and training variants on Qwen3-Reranker-4B.** Higher nDCG@10 and lower τ-PSI are better. RankZephyr and jina are listwise rerankers, whereas mxbai is pointwise and has zero τ-PSI by construction. The native jina row uses one context; the matched row uses the students' $B = 20$ geometry. Purpose-built systems use large hard-negative-mined reranking corpora, whereas OC-SFT uses about 30K MS MARCO queries, so they are external reference systems. Basis: *Public 16* excludes Legal-A and Legal-B.

Method	Public 16	DL19	DL20	DL21	DL22	DL23	Touché	FiQA	NFCorpus	ArguAna	Climate	T-COVID	DBPedia	SciFact	Signal1M	T-NEWS	Robust04	Legal-A	Legal-B
nDCG@10, higher is better																			
RankZephyr-7B	0.50	0.74	0.70	0.67	0.51	0.44	0.32	0.34	0.37	0.44	0.22	0.78	0.43	0.74	0.31	0.49	0.52	0.35	0.33
jina-reranker-v3	0.50	0.72	0.70	0.69	0.50	0.45	0.30	0.38	0.37	0.68	0.24	0.74	0.43	0.68	0.31	0.40	0.49	0.35	0.32
jina-reranker-v3 ($B = 20$)	0.47	0.69	0.65	0.63	0.48	0.42	0.25	0.35	0.34	0.65	0.23	0.73	0.39	0.64	0.25	0.35	0.45	0.27	0.26
mxbai-base-v2	0.50	0.73	0.67	0.66	0.51	0.43	0.35	0.36	0.35	0.59	0.31	0.70	0.44	0.67	0.33	0.42	0.49	0.31	0.27
mxbai-large-v2	0.52	0.73	0.69	0.69	0.50	0.46	0.33	0.39	0.36	0.67	0.35	0.74	0.45	0.73	0.31	0.47	0.51	0.33	0.32
Qwen3-Reranker-4B																			
Off the shelf	0.46	0.65	0.59	0.53	0.40	0.34	0.29	0.37	0.37	0.52	0.29	0.76	0.41	0.73	0.24	0.38	0.49	0.29	0.24
Order-avg.	0.51	0.75	0.69	0.66	0.49	0.46	0.30	0.40	0.39	0.53	0.29	0.79	0.46	0.73	0.30	0.44	0.53	0.37	0.32
OC-SFT	0.51	0.74	0.69	0.67	0.50	0.46	0.32	0.40	0.39	0.47	0.28	0.79	0.47	0.72	0.32	0.44	0.52	0.36	0.31
τ-PSI, lower is better																			
RankZephyr-7B	0.40	0.39	0.39	0.40	0.40	0.40	0.41	0.41	0.38	0.40	0.41	0.40	0.41	0.41	0.43	0.40	0.40	0.41	0.41
jina-reranker-v3	0.16	0.15	0.16	0.13	0.15	0.16	0.13	0.15	0.17	0.16	0.16	0.14	0.18	0.20	0.18	0.15	0.17	0.16	0.17
jina-reranker-v3 ($B = 20$)	0.18	0.17	0.17	0.16	0.18	0.18	0.15	0.17	0.18	0.18	0.18	0.15	0.20	0.20	0.20	0.17	0.18	0.18	0.18
Qwen3-Reranker-4B																			
Off the shelf	0.29	0.29	0.29	0.31	0.30	0.32	0.29	0.26	0.31	0.30	0.25	0.27	0.30	0.30	0.33	0.27	0.29	0.32	0.31
Order-avg.	0.12	0.11	0.11	0.10	0.11	0.12	0.10	0.10	0.15	0.14	0.11	0.09	0.14	0.14	0.16	0.11	0.13	0.13	0.12
OC-SFT	0.08	0.07	0.08	0.07	0.08	0.08	0.06	0.06	0.10	0.13	0.07	0.06	0.09	0.09	0.12	0.07	0.09	0.08	0.08

H.4 CROSS-BASE SUMMARIES

Across all twelve bases, OC-SFT trained from single-order labels is more stable than order-averaged distillation (Figures 9 and 10 and Table 36). OC-SFT trained from order-averaged labels reduces instability further on ten of the eleven dense bases and on the sparse MoE. Qwen3-0.6B is omitted because its off-the-shelf output is degenerate and does not follow the required format.

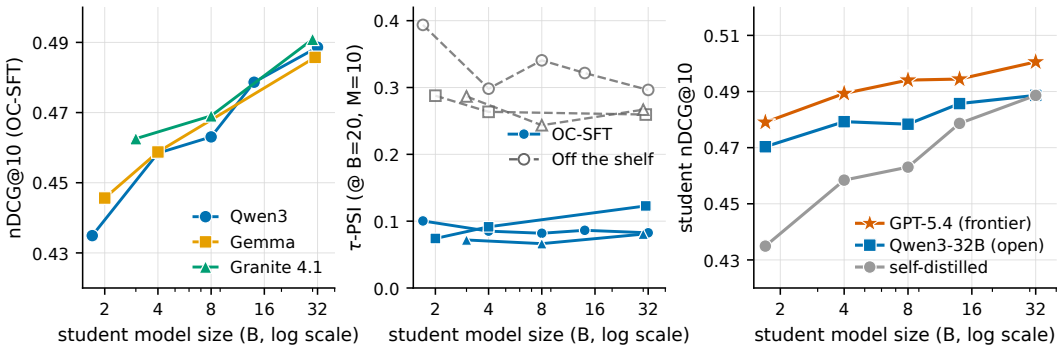


Figure 9: **Model size, family and teacher identity.** Panels (a) and (b) show mean nDCG@10 and τ-PSI over the 18 reranking collections, respectively, by model size for Qwen3, Gemma and Granite. Gemma E2B and E4B are plotted at their effective sizes, 2B and 4B. Panel (c) compares nDCG@10 across the evaluated Qwen3 student sizes for frontier, open and self-distilled teachers.

An open teacher approaches GPT-5.4 across student sizes (Figure 9, panel c). Using Qwen3-32B as the teacher keeps student nDCG@10 within 0.016 of GPT-5.4 and matches or exceeds self-distillation at every evaluated Qwen3 student size. Compared with self-distillation, the open teacher improves nDCG@10 by about 0.04 at Qwen3-1.7B, and the difference narrows as student size increases. GPT-5.4 remains highest at every size, but the open teacher closes most of that quality gap without using GPT-5.4 labels. Therefore, changing the teacher affects nDCG@10 more than τ-PSI.

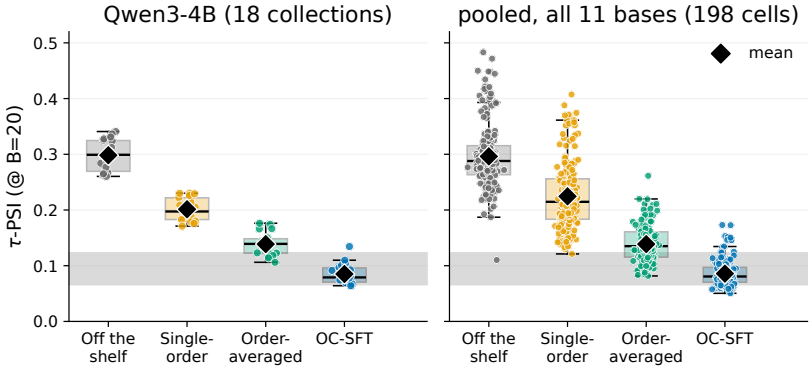


Figure 10: **Per-variant τ-PSI spread across collections.** Panel (a) shows Qwen3-4B across the 18 reranking collections. Panel (b) pools the same collections across all 11 dense bases, for 198 model-collection cells. Diamonds mark means; the shaded region spans the range of model-level mean OC-SFT τ-PSI values. The sparse MoE is excluded from this figure and reported in Table 36.

Teacher selection should consider label order-consistency as well as label quality (Figure 11). Silver-β measures how much a document's silver grade shifts when the teacher sees shuffled candidate orders; lower values denote more order-consistent labels. Formally, let $\mu_t(q,d)$ be a document's mean raw grade over $T{=}10$ teacher permutations and $\bar{\delta}_p$ the mean of $s^{(k)}_{q,d} - \mu_t(q,d)$ among its occurrences at slot p of the grading prompt; then

$$\text{silver-}\beta = \sqrt{\tfrac{1}{B}\textstyle\sum_{p=0}^{B-1} \bar{\delta}_p^{\,2}},$$

which is 0 when grades do not depend on order. It is a position-conditioned residual in grade units, not a rank correlation. Across the 8 teachers, silver-β and silver quality each correlate with student nDCG@10, but they also correlate with each other, so we cannot isolate either effect alone. The within-teacher control in App. D.3 varies label order-consistency while holding silver quality nearly fixed and produces a more stable student. Together, these results show why teacher evaluation should report both quality and order-consistency of its silver labels.

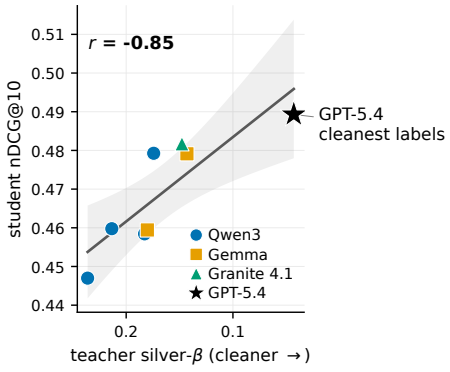


Figure 11: **Student quality correlates with label order-consistency across teachers.** Student nDCG@10 plotted against teacher silver-β for eight teachers. The correlation with student nDCG@10 is -0.852 for silver-β and $+0.812$ for silver quality; they correlate at -0.528. Shading marks the 95% confidence region for the fitted relationship between silver-β and student nDCG@10.

Table 35: **Per-collection nDCG@10 for every base and training variant.** Twelve bases, eleven dense and one sparse Mixture-of-Experts, each with its four training variants; higher is better. The final columns report means over the 16 public and all 18 reranking collections. Every block carries a fifth row, OC-SFT trained on order-averaged rather than single-order teacher labels. The first stage is BM25 on the 16 public collections. Legal-A and Legal-B use the precomputed candidate lists packaged with their internal fixtures; the retriever identity is not recorded. Basis: all 18 reranking collections, seed 42, one independently selected checkpoint per base and variant; off-the-shelf rows evaluate fixed base checkpoints. Order instability for the same cells is in Table 36.

Variant	DL19	DL20	DL21	DL22	DL23	Touché-2020	FiQA	NFCorpus	ArguAna	Climate-FEVER	TREC-COVID	DBPedia	SciFact	Signal-1M	TREC-NEWS	Robust04	Legal-A	Legal-B	Public 16	All 18
Qwen3-1.7B																				
Off the shelf	0.479	0.435	0.452	0.288	0.249	0.208	0.213	0.301	0.318	0.142	0.675	0.258	0.526	0.210	0.307	0.399	0.233	0.209	0.341	0.328
Single-order	0.586	0.547	0.545	0.347	0.317	0.246	0.215	0.289	0.262	0.165	0.703	0.315	0.525	0.232	0.352	0.430	0.180	0.174	0.380	0.357
Order-avg.	0.678	0.615	0.630	0.405	0.367	0.282	0.288	0.336	0.344	0.177	0.736	0.379	0.633	0.296	0.405	0.474	0.287	0.279	0.440	0.423
OC-SFT	0.695	0.633	0.662	0.427	0.391	0.290	0.296	0.347	0.286	0.202	0.728	0.417	0.638	0.292	0.432	0.495	0.305	0.291	0.452	0.435
OC-SFT (avg. labels)	0.678	0.622	0.646	0.421	0.375	0.279	0.301	0.349	0.306	0.190	0.751	0.393	0.657	0.296	0.417	0.477	0.277	0.260	0.447	0.427
Qwen3-4B																				
Off the shelf	0.596	0.552	0.542	0.389	0.331	0.221	0.251	0.312	0.206	0.157	0.724	0.334	0.512	0.217	0.330	0.450	0.286	0.243	0.383	0.370
Single-order	0.719	0.675	0.670	0.448	0.430	0.258	0.346	0.353	0.257	0.189	0.753	0.403	0.696	0.293	0.423	0.516	0.338	0.290	0.464	0.448
Order-avg.	0.727	0.675	0.674	0.453	0.437	0.286	0.349	0.364	0.283	0.208	0.747	0.413	0.708	0.294	0.443	0.520	0.359	0.310	0.474	0.458
OC-SFT	0.732	0.681	0.676	0.466	0.448	0.276	0.365	0.357	0.311	0.189	0.740	0.417	0.714	0.295	0.423	0.515	0.352	0.306	0.475	0.459
OC-SFT (avg. labels)	0.728	0.681	0.670	0.464	0.451	0.274	0.359	0.356	0.302	0.187	0.750	0.419	0.714	0.298	0.427	0.517	0.332	0.294	0.475	0.457
Qwen3-8B																				
Off the shelf	0.611	0.533	0.505	0.362	0.327	0.229	0.310	0.318	0.235	0.141	0.718	0.329	0.602	0.208	0.327	0.453	0.265	0.241	0.388	0.373
Single-order	0.705	0.659	0.640	0.444	0.424	0.261	0.357	0.341	0.325	0.162	0.738	0.397	0.705	0.271	0.403	0.507	0.320	0.261	0.459	0.440
Order-avg.	0.713	0.689	0.668	0.466	0.449	0.253	0.365	0.363	0.299	0.162	0.749	0.409	0.722	0.281	0.419	0.522	0.329	0.289	0.471	0.453
OC-SFT	0.729	0.702	0.683	0.482	0.462	0.278	0.380	0.360	0.320	0.178	0.750	0.416	0.723	0.282	0.418	0.528	0.353	0.291	0.481	0.463
OC-SFT (avg. labels)	0.727	0.687	0.673	0.477	0.448	0.262	0.374	0.362	0.335	0.166	0.762	0.413	0.725	0.282	0.421	0.523	0.353	0.278	0.477	0.459
Qwen3-14B																				
Off the shelf	0.642	0.587	0.587	0.405	0.370	0.214	0.357	0.338	0.340	0.143	0.748	0.342	0.712	0.228	0.354	0.491	0.278	0.237	0.429	0.410
Single-order	0.718	0.694	0.676	0.463	0.440	0.271	0.400	0.362	0.455	0.172	0.752	0.407	0.729	0.304	0.438	0.533	0.342	0.280	0.488	0.469
Order-avg.	0.733	0.702	0.699	0.482	0.447	0.260	0.399	0.375	0.445	0.156	0.780	0.415	0.744	0.298	0.465	0.553	0.339	0.286	0.497	0.477
OC-SFT	0.749	0.709	0.702	0.490	0.462	0.282	0.405	0.376	0.358	0.171	0.780	0.430	0.740	0.301	0.465	0.557	0.342	0.298	0.499	0.479
OC-SFT (avg. labels)	0.736	0.694	0.703	0.487	0.451	0.267	0.404	0.381	0.347	0.171	0.794	0.420	0.739	0.299	0.462	0.557	0.342	0.291	0.495	0.475
Qwen3-32B																				
Off the shelf	0.628	0.581	0.569	0.401	0.334	0.253	0.319	0.331	0.332	0.149	0.741	0.356	0.708	0.222	0.348	0.486	0.278	0.248	0.422	0.405
Single-order	0.706	0.674	0.660	0.469	0.425	0.259	0.392	0.342	0.430	0.158	0.741	0.407	0.731	0.283	0.444	0.530	0.329	0.284	0.478	0.459
Order-avg.	0.741	0.702	0.690	0.493	0.444	0.265	0.406	0.377	0.460	0.172	0.788	0.425	0.745	0.285	0.469	0.555	0.338	0.299	0.501	0.481
OC-SFT	0.748	0.712	0.694	0.507	0.460	0.276	0.417	0.380	0.464	0.177	0.776	0.435	0.743	0.288	0.493	0.570	0.350	0.302	0.509	0.488
OC-SFT (avg. labels)	0.742	0.702	0.694	0.502	0.451	0.267	0.413	0.385	0.449	0.174	0.783	0.431	0.739	0.290	0.467	0.560	0.338	0.293	0.503	0.482
Gemma-E2B																				
Off the shelf	0.586	0.518	0.581	0.365	0.312	0.224	0.235	0.296	0.296	0.143	0.699	0.316	0.526	0.213	0.307	0.443	0.280	0.225	0.379	0.365
Single-order	0.715	0.644	0.641	0.412	0.415	0.260	0.306	0.336	0.363	0.182	0.752	0.388	0.629	0.293	0.408	0.498	0.332	0.284	0.453	0.437
Order-avg.	0.709	0.644	0.656	0.433	0.429	0.255	0.311	0.338	0.404	0.186	0.748	0.385	0.654	0.309	0.397	0.496	0.355	0.301	0.460	0.445
OC-SFT	0.721	0.638	0.670	0.433	0.431	0.263	0.323	0.340	0.356	0.183	0.757	0.392	0.648	0.316	0.403	0.496	0.361	0.291	0.461	0.446

continued on the next page

Table 35, continued.

Variant	DL19	DL20	DL21	DL22	DL23	Touché-2020	FiQA	NFCorpus	ArguAna	Climate-FEVER	TREC-COVID	DBPedia	SciFact	Signal-1M	TREC-NEWS	Robust04	Legal-A	Legal-B	Public 16	All 18
OC-SFT (avg. labels)	0.727	0.634	0.659	0.442	0.427	0.277	0.317	0.340	0.358	0.185	0.751	0.400	0.661	0.311	0.417	0.502	0.356	0.293	0.463	0.448
Gemma-E4B																				
Off the shelf	0.632	0.593	0.616	0.403	0.369	0.215	0.303	0.327	0.221	0.124	0.747	0.343	0.595	0.219	0.353	0.499	0.305	0.240	0.410	0.395
Single-order	0.719	0.679	0.688	0.458	0.437	0.242	0.344	0.354	0.342	0.138	0.768	0.393	0.702	0.309	0.433	0.540	0.366	0.295	0.472	0.456
Order-avg.	0.727	0.680	0.688	0.472	0.446	0.256	0.353	0.364	0.392	0.152	0.764	0.404	0.709	0.310	0.448	0.541	0.362	0.298	0.482	0.465
OC-SFT	0.695	0.665	0.669	0.463	0.439	0.252	0.357	0.364	0.365	0.178	0.771	0.406	0.706	0.293	0.446	0.551	0.351	0.287	0.476	0.459
OC-SFT (avg. labels)	0.727	0.688	0.683	0.473	0.439	0.265	0.364	0.358	0.400	0.153	0.776	0.413	0.707	0.306	0.425	0.541	0.345	0.295	0.482	0.464
Gemma-31B																				
Off the shelf	0.642	0.632	0.624	0.465	0.416	0.242	0.413	0.355	0.413	0.163	0.756	0.368	0.720	0.226	0.371	0.543	0.354	0.281	0.459	0.444
Single-order	0.702	0.698	0.684	0.497	0.468	0.270	0.419	0.370	0.435	0.168	0.750	0.407	0.746	0.292	0.442	0.575	0.371	0.300	0.495	0.477
Order-avg.	0.722	0.711	0.706	0.506	0.480	0.277	0.420	0.377	0.464	0.191	0.751	0.419	0.741	0.311	0.463	0.572	0.386	0.319	0.507	0.490
OC-SFT	0.733	0.710	0.704	0.521	0.482	0.283	0.430	0.384	0.340	0.184	0.752	0.427	0.749	0.295	0.473	0.578	0.381	0.316	0.503	0.486
OC-SFT (avg. labels)	0.730	0.708	0.706	0.526	0.488	0.279	0.424	0.380	0.479	0.181	0.758	0.421	0.741	0.286	0.457	0.578	0.378	0.310	0.509	0.491
Granite-3B																				
Off the shelf	0.586	0.510	0.539	0.380	0.311	0.221	0.276	0.323	0.177	0.176	0.708	0.322	0.573	0.231	0.358	0.443	0.279	0.246	0.383	0.370
Single-order	0.715	0.667	0.665	0.449	0.419	0.273	0.319	0.347	0.255	0.211	0.753	0.407	0.667	0.321	0.430	0.506	0.320	0.283	0.463	0.445
Order-avg.	0.728	0.662	0.676	0.460	0.435	0.262	0.332	0.357	0.268	0.196	0.747	0.401	0.687	0.307	0.438	0.511	0.326	0.302	0.467	0.450
OC-SFT	0.738	0.677	0.680	0.464	0.428	0.288	0.338	0.360	0.349	0.219	0.756	0.422	0.693	0.326	0.445	0.527	0.317	0.298	0.482	0.462
OC-SFT (avg. labels)	0.731	0.665	0.676	0.464	0.422	0.263	0.339	0.365	0.317	0.208	0.759	0.412	0.692	0.322	0.440	0.520	0.320	0.302	0.475	0.456
Granite-8B																				
Off the shelf	0.637	0.587	0.584	0.402	0.376	0.240	0.325	0.354	0.277	0.153	0.752	0.364	0.683	0.240	0.397	0.483	0.296	0.258	0.428	0.412
Single-order	0.721	0.678	0.681	0.462	0.447	0.263	0.356	0.372	0.278	0.174	0.764	0.410	0.720	0.306	0.467	0.549	0.310	0.264	0.478	0.457
Order-avg.	0.720	0.675	0.682	0.477	0.454	0.264	0.355	0.374	0.309	0.175	0.764	0.414	0.719	0.301	0.468	0.552	0.334	0.295	0.481	0.463
OC-SFT	0.720	0.671	0.688	0.488	0.467	0.273	0.369	0.373	0.332	0.178	0.768	0.419	0.727	0.319	0.475	0.553	0.329	0.294	0.489	0.469
OC-SFT (avg. labels)	0.724	0.676	0.692	0.474	0.464	0.267	0.371	0.375	0.342	0.172	0.772	0.419	0.722	0.316	0.475	0.551	0.333	0.298	0.488	0.469
Granite-30B																				
Off the shelf	0.659	0.611	0.595	0.429	0.391	0.241	0.378	0.345	0.283	0.145	0.761	0.373	0.702	0.231	0.382	0.517	0.310	0.253	0.440	0.423
Single-order	0.722	0.681	0.673	0.488	0.457	0.271	0.388	0.361	0.393	0.155	0.766	0.428	0.730	0.299	0.455	0.551	0.351	0.300	0.489	0.471
Order-avg.	0.745	0.702	0.712	0.500	0.476	0.296	0.406	0.380	0.387	0.167	0.786	0.431	0.748	0.306	0.458	0.570	0.359	0.309	0.504	0.485
OC-SFT	0.746	0.717	0.707	0.510	0.480	0.304	0.413	0.380	0.410	0.180	0.782	0.444	0.747	0.311	0.479	0.574	0.349	0.300	0.511	0.491
OC-SFT (avg. labels)	0.745	0.710	0.711	0.503	0.473	0.311	0.410	0.383	0.439	0.173	0.774	0.438	0.740	0.312	0.474	0.570	0.342	0.296	0.510	0.489
Gemma-4 26B-A4B (sparse MoE)																				
Off the shelf	0.674	0.658	0.628	0.471	0.415	0.246	0.398	0.346	0.353	0.147	0.749	0.361	0.692	0.239	0.379	0.547	0.339	0.276	0.456	0.440
Single-order	0.715	0.702	0.682	0.503	0.461	0.264	0.414	0.369	0.434	0.169	0.752	0.405	0.741	0.300	0.443	0.569	0.379	0.304	0.495	0.478
Order-avg.	0.720	0.709	0.690	0.508	0.469	0.279	0.415	0.376	0.457	0.166	0.761	0.416	0.744	0.314	0.441	0.574	0.375	0.315	0.502	0.485
OC-SFT ($\lambda = 3$)	0.720	0.703	0.693	0.515	0.483	0.278	0.419	0.380	0.478	0.183	0.748	0.406	0.744	0.303	0.442	0.575	0.379	0.316	0.504	0.487
OC-SFT (avg. labels, $\lambda = 5$)	0.728	0.702	0.690	0.518	0.489	0.283	0.418	0.378	0.461	0.173	0.758	0.410	0.743	0.306	0.451	0.577	0.372	0.303	0.505	0.487

Table 36: **Per-collection τ-PSI for every base and training variant.** The same twelve bases and 18 per-collection cells as Table 35; lower is more stable. The *All 18* column averages all 18 reranking collections. Every block also includes OC-SFT trained on order-averaged labels. Basis: all 18 reranking collections, seed 42, $M = 10$ random permutations, one independently selected checkpoint per base and variant.

Variant	DL19	DL20	DL21	DL22	DL23	Touché-2020	FiQA	NFCorpus	ArguAna	Climate-FEVER	TREC-COVID	DBPedia	SciFact	Signal-1M	TREC-NEWS	Robust04	Legal-A	Legal-B	All 18
Qwen3-1.7B																			
Off the shelf	0.393	0.387	0.405	0.404	0.403	0.413	0.415	0.383	0.376	0.342	0.392	0.399	0.377	0.373	0.367	0.365	0.448	0.445	0.394
Single-order	0.329	0.324	0.336	0.352	0.352	0.315	0.361	0.355	0.365	0.388	0.337	0.361	0.407	0.357	0.334	0.370	0.371	0.374	0.355
Order-avg.	0.148	0.148	0.149	0.163	0.163	0.128	0.144	0.179	0.198	0.174	0.137	0.180	0.211	0.181	0.135	0.177	0.172	0.173	0.164
OC-SFT	0.094	0.099	0.084	0.097	0.105	0.074	0.081	0.101	0.153	0.096	0.079	0.104	0.110	0.132	0.087	0.105	0.106	0.101	0.100
OC-SFT (avg. labels)	0.065	0.067	0.062	0.071	0.071	0.057	0.063	0.080	0.114	0.077	0.062	0.085	0.081	0.096	0.064	0.082	0.072	0.071	0.074
Qwen3-4B																			
Off the shelf	0.302	0.296	0.312	0.325	0.328	0.288	0.263	0.284	0.267	0.265	0.277	0.324	0.305	0.331	0.265	0.260	0.341	0.336	0.298
Single-order	0.188	0.180	0.183	0.209	0.205	0.183	0.179	0.213	0.221	0.189	0.176	0.222	0.230	0.229	0.171	0.189	0.230	0.226	0.201
Order-avg.	0.123	0.123	0.116	0.139	0.137	0.114	0.118	0.150	0.174	0.143	0.106	0.166	0.176	0.166	0.123	0.141	0.142	0.139	0.139
OC-SFT	0.069	0.074	0.064	0.076	0.081	0.068	0.068	0.093	0.134	0.092	0.064	0.097	0.104	0.110	0.078	0.101	0.079	0.079	0.085
OC-SFT (avg. labels)	0.062	0.066	0.059	0.070	0.070	0.061	0.060	0.079	0.125	0.080	0.058	0.088	0.095	0.097	0.065	0.082	0.066	0.065	0.075
Qwen3-8B																			
Off the shelf	0.316	0.304	0.326	0.323	0.325	0.303	0.262	0.289	0.240	0.244	0.301	0.315	0.269	0.321	0.259	0.257	0.318	0.321	0.294
Single-order	0.256	0.240	0.240	0.257	0.252	0.246	0.253	0.296	0.267	0.282	0.240	0.288	0.316	0.278	0.249	0.268	0.270	0.270	0.265
Order-avg.	0.160	0.150	0.140	0.167	0.159	0.132	0.145	0.189	0.187	0.167	0.130	0.190	0.201	0.185	0.140	0.162	0.174	0.169	0.164
OC-SFT	0.073	0.076	0.067	0.078	0.079	0.062	0.065	0.090	0.125	0.085	0.060	0.097	0.103	0.108	0.069	0.091	0.075	0.072	0.082
OC-SFT (avg. labels)	0.065	0.066	0.062	0.068	0.070	0.058	0.056	0.075	0.106	0.070	0.056	0.084	0.080	0.089	0.061	0.076	0.064	0.063	0.071
Qwen3-14B																			
Off the shelf	0.284	0.280	0.286	0.288	0.297	0.301	0.238	0.272	0.212	0.238	0.267	0.296	0.255	0.309	0.245	0.240	0.285	0.294	0.272
Single-order	0.252	0.246	0.232	0.256	0.260	0.265	0.266	0.294	0.284	0.281	0.236	0.284	0.328	0.277	0.268	0.268	0.283	0.277	0.270
Order-avg.	0.161	0.159	0.137	0.167	0.168	0.147	0.164	0.199	0.179	0.209	0.126	0.219	0.261	0.192	0.175	0.188	0.179	0.167	0.178
OC-SFT	0.079	0.085	0.062	0.077	0.081	0.075	0.078	0.099	0.125	0.094	0.060	0.100	0.095	0.104	0.076	0.096	0.086	0.082	0.086
OC-SFT (avg. labels)	0.075	0.076	0.065	0.079	0.082	0.065	0.070	0.096	0.130	0.099	0.061	0.110	0.118	0.104	0.083	0.102	0.073	0.071	0.087
Qwen3-32B																			
Off the shelf	0.315	0.305	0.318	0.317	0.322	0.340	0.260	0.295	0.234	0.249	0.330	0.312	0.252	0.330	0.264	0.258	0.310	0.324	0.296
Single-order	0.248	0.228	0.229	0.242	0.241	0.285	0.234	0.318	0.217	0.249	0.249	0.277	0.290	0.253	0.235	0.265	0.258	0.265	0.255
Order-avg.	0.135	0.129	0.123	0.139	0.140	0.143	0.112	0.163	0.113	0.129	0.117	0.171	0.148	0.161	0.120	0.135	0.143	0.142	0.137
OC-SFT	0.083	0.085	0.073	0.086	0.089	0.082	0.064	0.080	0.070	0.079	0.066	0.101	0.095	0.108	0.076	0.086	0.082	0.080	0.083
OC-SFT (avg. labels)	0.063	0.066	0.059	0.068	0.071	0.059	0.054	0.073	0.069	0.069	0.053	0.090	0.075	0.092	0.060	0.074	0.062	0.059	0.068
Gemma-E2B																			
Off the shelf	0.301	0.306	0.298	0.308	0.311	0.269	0.268	0.249	0.279	0.248	0.273	0.301	0.276	0.324	0.263	0.237	0.335	0.332	0.288
Single-order	0.192	0.203	0.187	0.211	0.218	0.177	0.196	0.203	0.204	0.199	0.177	0.232	0.247	0.272	0.199	0.212	0.234	0.227	0.211
Order-avg.	0.116	0.127	0.111	0.131	0.134	0.104	0.111	0.127	0.113	0.119	0.102	0.154	0.134	0.167	0.115	0.123	0.143	0.136	0.126
OC-SFT	0.070	0.081	0.067	0.078	0.081	0.060	0.060	0.068	0.060	0.072	0.064	0.095	0.073	0.106	0.069	0.075	0.080	0.077	0.074
OC-SFT (avg. labels)	0.064	0.075	0.060	0.072	0.074	0.057	0.059	0.062	0.054	0.068	0.057	0.088	0.066	0.095	0.063	0.070	0.073	0.071	0.068
Gemma-E4B																			

continued on the next page

Table 36, continued.

Variant	DL19	DL20	DL21	DL22	DL23	Touché-2020	FiQA	NFCorpus	ArguAna	Climate-FEVER	TREC-COVID	DBPedia	SciFact	Signal-1M	TREC-NEWS	Robust04	Legal-A	Legal-B	All 18
Off the shelf	0.306	0.297	0.302	0.301	0.305	0.265	0.220	0.236	0.220	0.224	0.275	0.276	0.220	0.312	0.235	0.202	0.277	0.272	0.264
Single-order	0.185	0.179	0.175	0.197	0.198	0.150	0.141	0.169	0.158	0.169	0.149	0.200	0.186	0.213	0.144	0.138	0.208	0.199	0.175
Order-avg.	0.115	0.123	0.111	0.130	0.134	0.096	0.094	0.118	0.110	0.113	0.092	0.143	0.126	0.152	0.100	0.100	0.132	0.122	0.117
OC-SFT	0.087	0.097	0.081	0.095	0.099	0.070	0.071	0.094	0.097	0.089	0.065	0.117	0.112	0.118	0.080	0.092	0.096	0.090	0.092
OC-SFT (avg. labels)	0.072	0.078	0.068	0.077	0.080	0.060	0.057	0.076	0.056	0.069	0.056	0.094	0.077	0.093	0.063	0.069	0.075	0.068	0.072
Gemma-31B																			
Off the shelf	0.284	0.273	0.294	0.270	0.274	0.298	0.206	0.246	0.225	0.233	0.273	0.280	0.243	0.297	0.243	0.227	0.247	0.261	0.260
Single-order	0.222	0.200	0.210	0.216	0.220	0.226	0.178	0.236	0.192	0.209	0.189	0.247	0.249	0.246	0.193	0.208	0.220	0.221	0.216
Order-avg.	0.148	0.142	0.130	0.149	0.154	0.137	0.135	0.179	0.131	0.156	0.108	0.191	0.211	0.202	0.148	0.183	0.141	0.134	0.154
OC-SFT	0.112	0.109	0.095	0.110	0.116	0.093	0.108	0.146	0.173	0.123	0.079	0.152	0.173	0.150	0.114	0.146	0.111	0.102	0.123
OC-SFT (avg. labels)	0.087	0.087	0.077	0.090	0.095	0.071	0.070	0.094	0.071	0.092	0.063	0.121	0.113	0.129	0.084	0.098	0.085	0.078	0.089
Granite-3B																			
Off the shelf	0.291	0.285	0.290	0.300	0.299	0.264	0.301	0.272	0.277	0.268	0.267	0.292	0.314	0.296	0.262	0.272	0.303	0.306	0.287
Single-order	0.146	0.141	0.145	0.158	0.157	0.133	0.138	0.164	0.173	0.144	0.137	0.160	0.161	0.162	0.121	0.130	0.185	0.185	0.152
Order-avg.	0.098	0.100	0.092	0.106	0.108	0.086	0.095	0.116	0.138	0.106	0.086	0.116	0.123	0.127	0.091	0.103	0.114	0.117	0.107
OC-SFT	0.064	0.068	0.061	0.071	0.076	0.058	0.061	0.080	0.067	0.074	0.060	0.081	0.083	0.098	0.061	0.072	0.080	0.078	0.072
OC-SFT (avg. labels)	0.061	0.062	0.054	0.063	0.065	0.055	0.062	0.075	0.073	0.076	0.054	0.076	0.086	0.089	0.061	0.073	0.064	0.065	0.067
Granite-8B																			
Off the shelf	0.193	0.187	0.286	0.282	0.287	0.218	0.248	0.243	0.110	0.225	0.257	0.274	0.252	0.288	0.228	0.220	0.295	0.293	0.244
Single-order	0.147	0.146	0.139	0.148	0.154	0.150	0.142	0.175	0.194	0.169	0.135	0.182	0.195	0.182	0.143	0.163	0.186	0.186	0.163
Order-avg.	0.096	0.097	0.089	0.101	0.103	0.084	0.083	0.109	0.139	0.103	0.082	0.122	0.117	0.124	0.088	0.099	0.109	0.108	0.103
OC-SFT	0.058	0.063	0.055	0.064	0.068	0.054	0.053	0.069	0.085	0.067	0.050	0.082	0.075	0.085	0.059	0.075	0.067	0.065	0.066
OC-SFT (avg. labels)	0.054	0.058	0.052	0.059	0.062	0.049	0.048	0.063	0.068	0.061	0.048	0.077	0.067	0.081	0.053	0.064	0.059	0.059	0.060
Granite-30B																			
Off the shelf	0.274	0.271	0.273	0.276	0.283	0.276	0.248	0.249	0.235	0.236	0.270	0.288	0.271	0.300	0.239	0.241	0.282	0.289	0.267
Single-order	0.210	0.195	0.189	0.194	0.191	0.225	0.189	0.235	0.169	0.212	0.195	0.237	0.253	0.202	0.194	0.207	0.210	0.219	0.207
Order-avg.	0.128	0.125	0.110	0.128	0.128	0.115	0.112	0.153	0.120	0.157	0.100	0.173	0.220	0.155	0.141	0.153	0.130	0.127	0.137
OC-SFT	0.076	0.083	0.065	0.076	0.080	0.065	0.062	0.091	0.073	0.087	0.058	0.103	0.115	0.103	0.076	0.096	0.072	0.069	0.081
OC-SFT (avg. labels)	0.066	0.071	0.060	0.068	0.069	0.057	0.056	0.081	0.059	0.085	0.054	0.096	0.101	0.094	0.069	0.086	0.061	0.060	0.072
Gemma-4 26B-A4B (sparse MoE)																			
Off the shelf	0.267	0.251	0.277	0.254	0.263	0.288	0.193	0.233	0.197	0.207	0.256	0.268	0.225	0.279	0.219	0.201	0.249	0.268	0.244
Single-order	0.191	0.176	0.171	0.183	0.190	0.193	0.157	0.215	0.158	0.184	0.146	0.228	0.229	0.230	0.170	0.185	0.186	0.186	0.188
Order-avg.	0.134	0.131	0.117	0.134	0.139	0.134	0.108	0.152	0.120	0.129	0.097	0.176	0.156	0.171	0.123	0.134	0.132	0.128	0.134
OC-SFT ($\lambda = 3$)	0.113	0.114	0.096	0.112	0.120	0.091	0.090	0.129	0.098	0.120	0.076	0.161	0.154	0.154	0.109	0.125	0.106	0.101	0.115
OC-SFT (avg. labels, $\lambda = 5$)	0.088	0.087	0.076	0.087	0.095	0.073	0.066	0.092	0.066	0.084	0.060	0.120	0.098	0.114	0.076	0.085	0.081	0.076	0.085