RUBRIC-CALIBRATED PREFERENCES: CROSS-QUERY CALIBRATION OF LLM JUDGMENTS VIA ITEM RESPONSE THEORY

Anonymous authors
Paper under double-blind review

ABSTRACT

Rerankers decide which documents users and LLMs see, yet their standard metric, nDCG, relies on human relevance labels that are costly, sparse, noisy, and discretely graded. As rerankers approach each other in quality, nDCG on these labels therefore increasingly fails to separate them. LLM judges could supply dense labels. Relative judgments within one query tell even close candidates apart, yet their scores share no scale across queries. Absolute grades share one scale but are too coarse to distinguish documents of similar relevance. In this work, we propose Rubric-Calibrated Preferences (RCP), which combine both kinds of judgment. A listwise Bradley–Terry tournament orders each query's documents, and a rubric of yes/no criteria of increasing stringency provides an absolute standard. Item Response Theory (IRT), which scores test-takers based on their answers to common questions, then uses the shared criteria to put all queries' tournament scores on one scale. RCP's retrieval metric, RCP-nDCG, replaces nDCG's discrete labels with the resulting calibrated relevance probabilities. Against blind grades from 46 external annotators, calibration raises the correlation between a query's mean score and its mean human grade from 0.538 to 0.795. The probabilities rank a useful document above a non-useful one with probability 0.910 (AUC, chance 0.5), versus 0.651 for the benchmark labels. When the annotators' grades prefer one of two rerankers and exactly one metric agrees, that metric is RCP-nDCG in 72.4% of 185 comparisons (chance about 53%). On TREC-DL, RCP-nDCG sides with NIST assessors' grades on every reranker pair that these grades separate significantly. RCP-nDCG also resolves many of nDCG's ties and separates 1.9 times as many reranker pairs on NanoBEIR. Rubric calibration thus turns relative LLM judgments into dense relevance labels that are comparable across queries and agree with human judgment.

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.

1 INTRODUCTION

Reranking is central to modern retrieval. First-stage retrievers maximize recall, but rerankers select the top documents (Nogueira & Cho, 2019; Khattab & Zaharia, 2020, *inter alia*). Yet the standard reranking metric, nDCG@k, is increasingly inadequate. Per query, it saturates when all rerankers score 1, floors when all score 0, and compresses them when half or more of their pairs tie (Figure 1). For the 14 rerankers we evaluate, one of these failure modes occurs on 45.0% of NanoBEIR and 30.7% of BRIGHT queries. These failures stem from human relevance labels (qrels) that are costly, sparse, noisy, and discretely graded (Zobel, 1998; Buckley & Voorhees, 2004; Ban et al., 2026), and worsen as rerankers improve (Arabzadeh et al., 2022).

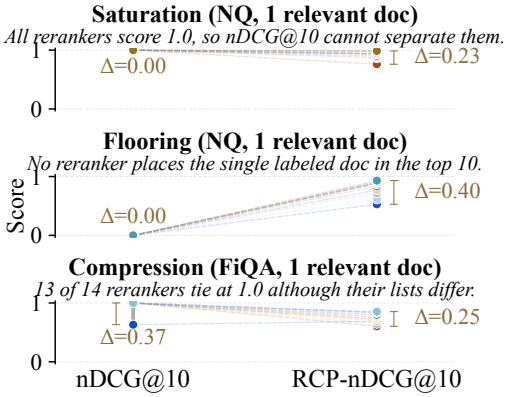


Figure 1: On three queries, nDCG@10 saturates, floors, or compresses 14 rerankers, and RCP-nDCG@10 spreads them.

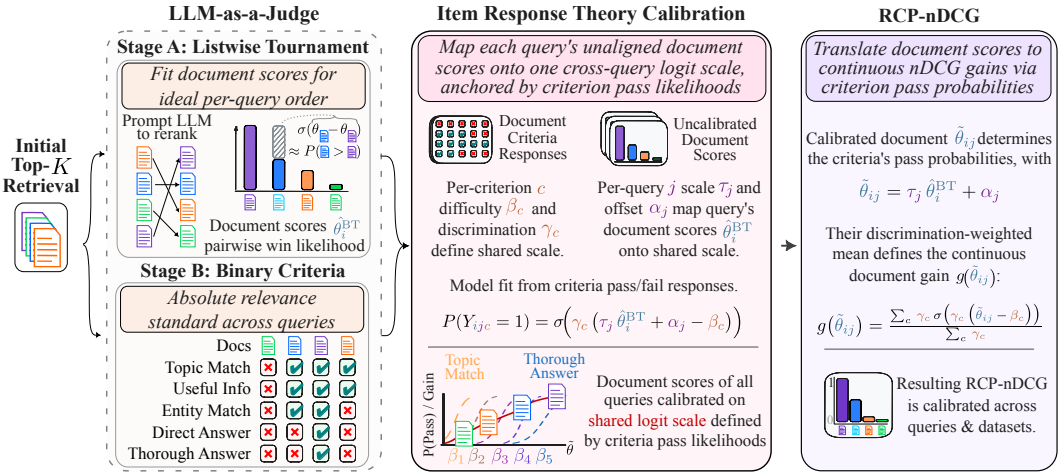


Figure 2: RCP(-nDCG). For each candidate document, **Stage A** produces per-query Bradley–Terry scores, and **Stage B** records criterion answers. An Item Response Theory **Calibration** model places the scores of all queries on one scale, which yields the continuous gains of RCP-nDCG (Section 3).

Large language models (LLMs) offer a promising alternative (Faggioli et al., 2023), generating relevance judgments at scale and at a quality competitive with human assessments (Thomas et al., 2024; Rahmani et al., 2025). However, relative judgments such as pair- or listwise comparisons (Sun et al., 2023; Yoon et al., 2025) produce fine-grained within-query rankings but no common scale across queries, which rules out graded labels, shared relevance thresholds, and cross-query statistics. Absolute judgments (Zhuang et al., 2024; Long et al., 2025) enable cross-query comparison, but their coarse grades leave many documents of a query tied. What is missing is a method that combines the resolution of relative judgments with the cross-query comparability of absolute ones. Our research question is whether a fixed rubric of yes/no criteria can place LLM preferences made within one query on a scale shared across queries. We also test whether the resulting scores agree with human judgment. We study this question in reranking evaluation, where independent human references exist and such scores could replace sparse, discrete, and noisy qrels.

We propose Rubric-Calibrated Preferences (RCP; Figure 2), a calibration method grounded in Item Response Theory (IRT). IRT scores test-takers on one scale based on their answers to shared questions (Lord, 1980). In RCP, documents take the test, and the criteria of a rubric are the questions. In Stage A, a listwise Bradley–Terry tournament (Bradley & Terry, 1952) over windows of documents produces per-document scores for a fine-grained within-query ordering. In Stage B, each document is evaluated against five binary criteria of increasing stringency, yielding an absolute, interpretable relevance signal. An IRT model (Birnbaum, 1968) then uses the shared criteria to place the tournament scores of all queries on one scale. Stage A sets the order within a query; the rubric sets the scale across queries. In retrieval, RCP labels a benchmark's candidate pool once, and RCP-nDCG then scores any reranker with the resulting dense gains instead of discrete labels.

Our contributions are fourfold. **(1)** We propose RCP, a calibration method that uses IRT to place within-query LLM preferences on a scale shared across queries. **(2)** We derive RCP-nDCG, a continuous retrieval metric whose gains are calibrated relevance probabilities rather than discrete labels. We evaluate it on 13 NanoBEIR datasets (Zeta Alpha, 2024; Muennighoff et al., 2023), 12 reasoning-intensive BRIGHT tasks (Su et al., 2025), and the eight corpora of ViDoRe v3 (Loison et al., 2026), with 14 rerankers from six model families. On NanoBEIR, RCP-nDCG nearly doubles the share of reranker pairs that a paired t-test separates significantly, from 34.2% for nDCG to 63.2% (Section 4.4). **(3)** We test whether these separations are correct against two human references. First, 46 external annotators graded the top-five lists of reranker pairs without seeing labels, scores, or which list a document came from. Against their grades, the AUC[1] of the RCP gains is 0.910, versus 0.651 for the qrels (Section 4.3). Where exactly one of nDCG and RCP-nDCG favors the annotators' preferred reranker, it is RCP-nDCG in 72.4% of 185 contests (chance about 53%, as nDCG ties more often; Section 4.5). Second, the official grades of the TREC Deep Learning track

[1]AUC: the probability that a label source ranks a random document that annotators rate useful above a random one they rate not useful (0.5 is chance, 1 is perfect).

(TREC-DL) (Craswell et al., 2020; 2021) come from NIST assessors. With either of two LLM judges, RCP-nDCG agrees with NIST on every reranker pair that these grades separate significantly. **(4)** We provide the code and data in the supplementary material and will release them publicly. Without further judge calls, the calibrated labels score any retrieval model that reorders our candidate pools. All LLM judgments and the 7,080 human grades can benchmark future LLM relevance judges.

2 Background

Notation and nDCG. We index queries q_j by j, a query's K candidate documents d_i by i, and rubric criteria by $c = 1, \ldots, C$. For a ranking π with $d_{\pi(r)}$ at rank r, the discounted cumulative gain (DCG) and its normalized form, nDCG (Järvelin & Kekäläinen, 2002), at cutoff k are

$$\mathrm{DCG@}k(\pi) = \sum_{r=1}^{k} G\big(d_{\pi(r)}\big)/\log_2(r+1), \quad \mathrm{nDCG@}k(\pi) = \mathrm{DCG@}k(\pi)/\mathrm{DCG@}k(\pi^*). \tag{1}$$

The ideal ranking π^* sorts all of the query's documents by decreasing gain. Benchmarks use qrels as the gains $G(d)$, give unlabeled documents gain 0, and average nDCG@k over queries.

Bradley–Terry model. The Bradley–Terry (BT) model (Bradley & Terry, 1952) gives the probability that document i is preferred to document i' as $P(i \succ i') = \sigma(\theta_i - \theta_{i'})$, where θ_i and $\theta_{i'}$ are latent scores and $\sigma(x) = 1/(1+e^{-x})$. Within a comparison set $\mathcal{I}$, θ_i is identified only up to an additive constant. Scores from different sets $\mathcal{I} \neq \mathcal{I}'$, such as two queries, are therefore not comparable. Their spread can also differ, because a judge may be more decisive in one set than in another.

Item Response Theory. In the two-parameter logistic (2PL) model of IRT (Birnbaum, 1968), a respondent with ability θ_i, here document i, answers an item, here criterion c, positively with probability $P(Y_{ic} = 1) = \sigma\big(\gamma_c(\theta_i - \beta_c)\big)$. The difficulty β_c is the ability at which criterion c is passed with probability one half, and the discrimination $\gamma_c > 0$ sets how steeply this pass probability rises with ability. Because the documents of all queries share the criterion parameters, equal abilities imply equal pass probabilities. All abilities thus share one scale. With only C yes/no criteria, many documents receive identical answers and ability estimates. IRT alone thus cannot order them finely.

3 Methodology

3.1 Stage A: Listwise Bradley–Terry Tournament

Given a query q_j and its candidate set $\mathcal{D}_j = \{d_1, \ldots, d_K\}$, Stage A repeats listwise LLM reranking (Sun et al., 2023) over windows of $w = 10$ documents. For each window, the LLM judge assigns every document a relevance score $s_i \in [-5, +5]$, which the prompt describes as a logit scale (Appendix J). Only the $w(w-1)/2$ soft preferences $p_{ii'} = \sigma(s_i - s_{i'})$ are kept. BT scores $\hat{\theta}_i^{\mathrm{BT}}$ are fit to these preferences (Pipitone et al., 2025) by minimizing the pairwise cross-entropy between $p_{ii'}$ and the predicted preferences $\sigma(\hat{\theta}_i^{\mathrm{BT}} - \hat{\theta}_{i'}^{\mathrm{BT}})$. A three-phase window schedule focuses on uncertain ranks and keeps mid-ranked documents away from dominant ones, which would lower their scores (Appendix C.2). Stage A ranks the documents by $\hat{\theta}_i^{\mathrm{BT}}$.

3.2 Stage B: Pointwise Criterion Evaluation

The LLM judge reads w documents per window but rates each document on its own against an absolute standard (Appendix J). Windows are drawn first at random, then among documents of similar estimated quality. Each document appears in n_{ij} Stage B windows, and each appearance yields binary responses $Y_{ijc} \in \{0, 1\}$ recording whether d_i passes criterion c of Table 1 for query q_j.

Rubric design. The five criteria target general-purpose retrieval, and other applications may need other criteria. In an ablation on TREC-DL, we test the shipped rubric and fifteen variants (Appendix G.6). No significantly separated reranker pair reverses unless the rubric ignores the query or puts all criteria at one end of the scale.

Table 1: Stage B criteria with discrimination γ_c and difficulty β_c on NanoBEIR (QWEN3.5-397B).

ID	Criterion	Diagnostic Question	γ_c	β_c
C1	Topical Relevance	*Is its topic clearly related to the query's information need?*	1.17	−2.57
C2	Information Utility	*Does it contain specific, useful information to address the query?*	1.05	−2.00
C3	Entity/Detail Match	*Does it name and discuss the particular entity or detail in the query?*	0.94	−0.25
C4	Direct Answer	*Does it explicitly and directly answer the query's primary question?*	1.16	0.71
C5	Thorough Treatment	*Does it treat the topic in depth rather than superficially?*	0.69	4.12

3.3 CALIBRATION: 2PL IRT MERGE

Since BT scores and 2PL abilities both lie on logit scales, we map BT scores affinely to 2PL abilities:

$$P(Y_{ijc} = 1) = \sigma\Big(\gamma_c \left(\tau_j \, \hat{\theta}_i^{\mathrm{BT}} + \alpha_j - \beta_c\right)\Big). \tag{2}$$

Each criterion's discrimination γ_c and difficulty β_c are shared across queries. The offset α_j places query j's BT scores on the shared scale, fixing BT's free additive constant (Section 2). The scale factor $\tau_j > 0$ sets their spread. The same BT gap can stand for a large relevance difference on one query and a small one on another. Scaling by τ_j gives a one-logit difference in ability the same relevance meaning on every query (Section 4.1). The Stage B answers of the query's documents estimate τ_j and α_j. Shifting or stretching all abilities, with matching changes to β_c and γ_c, leaves every probability in Equation (2) unchanged. We therefore fix the unit with $\sum_c \gamma_c = C$ and the origin with $\mathrm{mean}_c\, \beta_c = 0$ (Appendix I.2). With $\hat{\theta}_i^{\mathrm{BT}}$ held fixed, we fit all other parameters by maximizing their posterior under Equation (2) with weak priors, once per suite and judge (Appendix C.3). The calibrated ability of document d_i on query q_j is

$$\tilde{\theta}_{ij} = \tau_j \, \hat{\theta}_i^{\mathrm{BT}} + \alpha_j. \tag{3}$$

Proposition 3.1 (Order preservation)**.** *For any rubric and fit, sorting a query's documents by $\tilde{\theta}_{ij}$, or by any strictly increasing function of it, reproduces their Stage A order (Appendix I.1).*

Calibration serves to make scores comparable across queries: a document with $\tilde{\theta}_{ij} = \beta_c$ passes criterion c with probability one half on every query. Relevance labels, thresholds, and pooled statistics require this comparability (Section 4.2).

3.4 RCP-NDCG: A CONTINUOUS RELEVANCE METRIC

RCP-nDCG replaces the qrel gains of Equation (1) with a continuous gain:

$$g(\tilde{\theta}_{ij}) = \sum\nolimits_c \gamma_c \, \sigma\big(\gamma_c \left(\tilde{\theta}_{ij} - \beta_c\right)\big) \Big/ \sum\nolimits_c \gamma_c. \tag{4}$$

Weighting each criterion's pass probability by its discrimination makes g rise fastest where the criteria measure most precisely: its slope is proportional to the rubric's Fisher information. For any rubric, g rises smoothly from 0 to 1 (Lemma I.1). RCP-nDCG@k is nDCG@k with gain $G(d_i) = g(\tilde{\theta}_{ij})$ instead of qrels, and its ideal ranking is the Stage A order (Proposition 3.1). On NanoBEIR, g is 0.22 at C1's difficulty and 0.69 at C4's. In Figure 3, 79.3% of pool documents lie below C1's, 15.6% in between, and 5.1% above C4's. The gain thus varies most above C1's.

Count-nDCG. Let d_i pass criterion c in S_{ijc} of its n_{ij} Stage B windows for query q_j. Our baseline Count-nDCG uses the pass share as gain, like rubric-based judges (Farzi & Dietz, 2025),

$$G_{\mathrm{Count}}(d_i) = \sum\nolimits_c S_{ijc} / \left(C \, n_{ij}\right). \tag{5}$$

Documents with equal pass shares therefore tie. With equal discriminations and the ability implied by the Stage B answers alone, RCP-nDCG reduces to Count-nDCG (Proposition I.3).

4 EXPERIMENTS

Our experiments ask three questions. Is the 2PL well fitted, and are RCP's scores calibrated across queries (Sections 4.1 and 4.2)? Do they label documents better than the qrels (Section 4.3)? Does RCP-nDCG compare rerankers better than nDCG on the benchmarks' qrels, which we call qrel-nDCG (Sections 4.4 and 4.5)?

Datasets, pool, and rerankers. We evaluate on 13 NanoBEIR datasets (Zeta Alpha, 2024; Muennighoff et al., 2023) (649 queries) and 12 BRIGHT tasks (Su et al., 2025) (1,366 queries), two English-only suites with binary qrels. As in BRIGHT's official evaluation, we remove each query's excluded documents from every scored ranking (Appendix C.5). We also use TREC Deep Learning 2019 and 2020 (TREC-DL; 43 and 54 queries), whose MS MARCO passages NIST assessors graded from 0 to 3 (Nguyen et al., 2016; Craswell et al., 2020; 2021). The eight corpora of ViDoRe v3 (Loison et al., 2026) (2,419 questions) hold pages of enterprise, governmental, and educational documents that human annotators graded on a three-point scale. The judge and the rerankers read each page as the benchmark's OCR text (Appendix C.6). We produce the top-150 candidates per query ($K = 150$) via Reciprocal Rank Fusion (Cormack et al., 2009) of three retrievers: Cohere Embed v4, Octen-Embedding-8B, and BM25. Positive qrels that the retrievers miss replace the lowest-ranked candidates, to isolate reranking from first-stage recall. On TREC-DL, each pool joins these candidates with every passage NIST graded for the query (287 passages on average). The rerankers reorder the whole pool (Appendix C.7). We evaluate 14 rerankers from six families (Cohere, Contextual AI, Qwen3 Embedding, Voyage AI, Jina AI, ZeroEntropy; Appendix C.1).

Judges. Both stages use open-weight LLMs, chosen for cost and reproducibility. The primary judge is QWEN3.5-397B-NVFP4 (17B active) (NVIDIA, 2026). As a second judge, GPT-OSS-120B (5.1B active) (OpenAI, 2025) repeats both stages on NanoBEIR, BRIGHT, and TREC-DL (Appendix G.4). On TREC-DL, the smaller QWEN3.6-27B (Qwen Team, 2026b) replaces the primary judge (Appendix C.7).

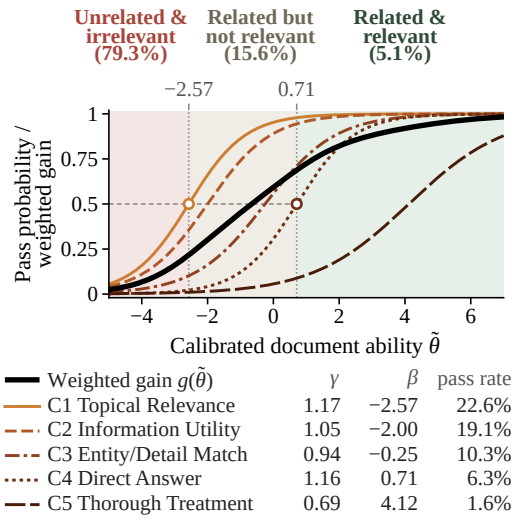


Figure 3: Criterion pass probabilities and gain g (NanoBEIR, QWEN3.5-397B). The legend gives discrimination γ, difficulty β, and pass rate. Shading splits at C1's and C4's difficulties.

Human annotations. We collected blind relevance grades from 46 paid external contractors, each with at least a BSc in the annotated subject (Appendix D). They graded documents from 12 NanoBEIR datasets, 7 BRIGHT tasks, and the five English corpora of ViDoRe v3. Each contest pairs a query with two rerankers' top-five lists. Annotators grade the union of both lists in random order, without seeing list membership, qrels, or scores. Per document, they answer four questions (Topical, Useful, Answers, Complete), and the most stringent yes sets the grade from 1 to 4, or 0 if none. Answers and Complete also allow Either, which gives a half grade. Most contests are disagreements, on which RCP-nDCG@10 and qrel-nDCG@10 favor different rerankers over the full rankings, and the rest serve as concordant controls (Appendix D.1). We analyze 311 contests on 292 distinct queries, each contest reviewed by three different annotators. Some queries enter two contests with different reranker pairs. A document's label is its mean grade from at least three annotators (7,080 grades, 2,268 documents; Appendix D.4). For each annotator, we score both rerankers' top-five lists by nDCG@5, with gains from that annotator's grades (Appendix D.2). The reranker with the higher score is the annotator's preferred one. A score gap below 0.005 is a tie, which the annotator's rankings of equally graded documents can break. The contest's verdict is the reranker that more of the three annotators prefer. In 22 contests, both are preferred equally often (one each and a tie, or three ties), leaving no verdict.

4.1 2PL DIAGNOSTICS

We first test how well the 2PL fits the judge's answers to the criteria, since calibration places all queries on one scale only if this fit holds. On NanoBEIR with QWEN3.5-397B, difficulty increases from C1 to C5. C1 to C4 have near-uniform discriminations, separating documents about equally sharply (Table 1 and Figure 3). The per-query scale τ_j varies about threefold, from 0.86 to 2.72 between its 5th and 95th percentiles, and similarly for GPT-OSS-120B and on BRIGHT (Appendix E.1). The 2PL's independence assumption is met: once the calibrated score is accounted for, two criteria's

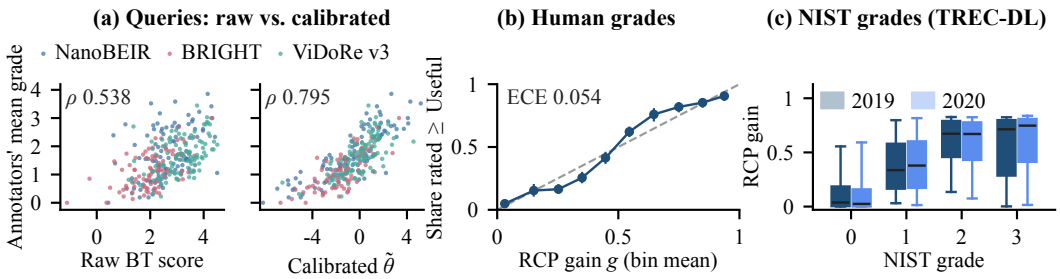


Figure 4: Calibration of RCP against human and NIST grades. (a) Per human-study query: annotators' mean grade against mean raw BT score (left) and mean calibrated score (right), with Spearman ρ. (b) Documents binned by gain g: share of their individual grades at Useful or better against the bin's mean gain (diagonal: perfect calibration). (c) RCP gain of TREC-DL passages by NIST grade.

answers correlate only weakly (mean absolute correlation 0.123, against 0.60 for raw answers; Appendix E.2). Criterion parameters also transfer across datasets. When fitted on all other datasets, they cut the cross-entropy of a held-out dataset's answers by 60.6% against per-criterion pass rates (Appendix E.1). The 2PL thus fits the judge's answers well enough to place all queries on one scale.

4.2 Calibration Quality

We then test the calibrated scores against three references: the judge's own answers, our human grades, and NIST's grades.

Judge's own passes. As RCP-nDCG's gains derive from predicted pass probabilities, we check them against the judge's 486,750 (query, document, criterion) answers on NanoBEIR (Appendix E.3). The Expected Calibration Error (ECE) averages the gap between predicted probability and observed pass rate over weighted bins (0 is perfect). A sigmoid of the raw BT score, ignoring each query's scale, is miscalibrated (ECE = 0.371), while the 2PL is well calibrated (ECE = 0.017).

Human grades. Calibration also aligns the gain with blind human grades (Figure 4b). If g is the probability that one annotator rates a document Useful or better, about 30% of the grades of documents with $g \approx 0.3$ should be Useful or better. Figure 4b bins documents by g and plots this share against each bin's mean gain. Against our 7,080 human grades, the ECE is 0.054 (95% CI [0.042, 0.071]; Appendix E.3). Calibration keeps each query's order and changes only how documents of different queries compare. Across queries, the Spearman correlation of a query's mean score with its annotators' mean grade rises from 0.538 for raw BT to 0.795 after calibration (Figure 4a).

NIST grades. On TREC-DL, NIST's graded judgments share no annotations with our study. Within a query, the share of differently graded passage pairs that RCP orders as NIST does is 0.874 (2019) and 0.828 (2020) (0.5 is chance). This share reflects Stage A alone (Proposition 3.1). Over the graded passages of all queries, the Spearman correlation of the calibrated score $\tilde{\theta}$ with the NIST grade is 0.649 and 0.595. With each query's scores in NIST's grade order, these correlations would be 0.768 and 0.743. The observed values thus reach 84.4% and 80.2% of this ceiling. In 2019, the spread of gains at grade 3 (Figure 4c) comes partly from positives that re-assessors rate low (Appendix H.2).

Calibration across queries. In every suite, calibration raises the Spearman correlation between a document's score and its mean grade, pooled over all queries, from 0.641–0.693 for raw BT to 0.753–0.830 (95% CIs of the gains above zero; Appendix E.3). The calibrated score tracks that grade as closely as a second, independent panel of annotators (Spearman 0.792 against 0.773; Appendix F.1).

4.3 Blind Evaluation of the Labels

A metric is only as good as its relevance labels. We therefore test RCP's labels against the qrels, first on whole sets of documents, then on single documents. On 264 NanoBEIR queries, the set of positive qrel documents differs from the equally large set with the highest RCP gains. By majority,

three blind LLM judges from other model families prefer the RCP set on 77.9% of the 222 queries they decide, and on 87.8% on BRIGHT (Appendix F.2). Where two IR experts, judging blind, agree (134 queries), they prefer the RCP set in 89.6%.

The gains also label single documents better than the qrels. Our human study grades the 2,268 documents that two rerankers place in their top five, blind to qrels and scores. We draw one of these documents that annotators rate useful (mean grade at least 2) and one they do not. The RCP gain ranks the useful one higher with probability 0.910 (AUC, ties counting half, chance 0.5). The binary qrels do so with probability 0.651, because they give about half of such pairs the same label. Of these documents, annotators rate 27.8% of those without a positive qrel as useful and 37.5% of those with one as not useful (Appendix F.1). The qrels thus disagree with the annotators in both directions.

4.4 RCP-nDCG Ranking Results

Reranker comparison. The largest change from qrel-nDCG@10 to RCP-nDCG@10 on NanoBEIR is the rise of the ZeroEntropy rerankers (Figure 5). Under qrel-nDCG@10, Contextual v2 6B and Qwen3-Reranker-8B lead with 74.1 and 74.0, and the three ZeroEntropy models trail (66.0–66.5). Under RCP-nDCG@10, zerank-2 and zerank-1 move from ranks 14 and 13 to ranks 1 and 2 (86.9, 86.8), and the two former leaders fall to ranks 6 and 9.

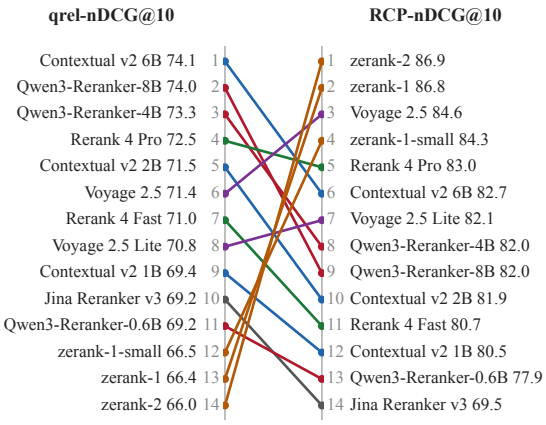


Figure 5: 14 rerankers on NanoBEIR under qrel-nDCG@10 (left) and RCP-nDCG@10 (right, Qwen3.5-397B), with means over datasets, colored by model family. Per-dataset values: Table 11.

Does the judge favor LLM-trained rerankers? The zerank rerankers train on pairwise preferences of an LLM ensemble (Pipitone et al., 2025). A metric built on an LLM judge could thus reward this training signal rather than quality. On TREC-DL, where NIST's grades test this concern, RCP-nDCG@10 (judge Qwen3.6-27B) ranks zerank-2 first of 18 systems (14 rerankers and four first-stage rankings) in both years. Under qrel-nDCG@10, which uses the NIST grades, zerank-2 ranks 14th in 2019 and 9th in 2020, but its gap to each year's leader is not significant (the 95% CI includes zero). Part of this rank deficit arises because NIST graded only passages retrieved by the systems submitted to the 2019 and 2020 tracks. Our pools add candidates from current retrievers, and nDCG counts every ungraded passage as irrelevant. Such passages fill 8.9% of zerank-2's top-10 slots, against 3.7–5.4% for the 9 systems from other families that NIST ranks above it with both years pooled. Our human study on NanoBEIR, BRIGHT, and ViDoRe v3 grades without any LLM, and annotators never see system names. In 66 contests with a verdict, RCP-nDCG favors the zerank reranker and qrel-nDCG the other. The annotators prefer zerank in 78.8% of them (Appendix G.5). NIST's grades thus do not contradict zerank's rise, which our annotators support.

RCP-nDCG is more discriminative. A metric separates two rerankers on a dataset if a paired t-test on their per-query scores rejects equal means at level 0.05. Its sensitivity (Clarke et al., 2021) is the share of the 91 pairs of our 14 rerankers that it separates, averaged over a suite's datasets.[2] On NanoBEIR, sensitivity rises from 34.2% under qrel-nDCG@10 to 63.2% under RCP-nDCG@10 (Figure 6a). RCP-nDCG@10 also separates more pairs on BRIGHT, on ViDoRe v3, and in both TREC-DL years, where qrel-nDCG@10 uses the NIST grades. The increase holds on most NanoBEIR datasets, on every BRIGHT task, and with gpt-oss-120b as judge (Appendices G.2 and G.4). Sensitivity measures separation, not correctness, which Section 4.5 tests.

RCP-nDCG vs. Count-nDCG. Count-nDCG is the rubric-only variant: a document's gain is its share of passed criteria in Stage B, without the tournament (Equation (5)). Count-nDCG separates

[2]If no two rerankers differed, the test would still separate about 5% of pairs by chance.

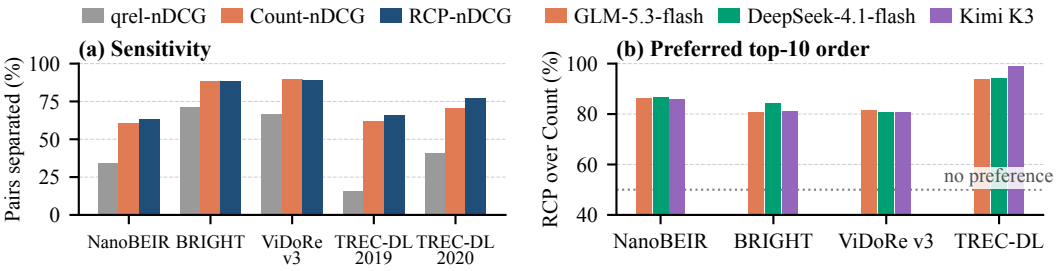


Figure 6: (a) Sensitivity: share of the 91 reranker pairs that each metric separates (paired t-test), mean over datasets. (b) Share of queries on which each judge prefers RCP-nDCG's ideal top-10 over the rubric-only Count-nDCG's, among queries where it prefers either list (50%: no preference).

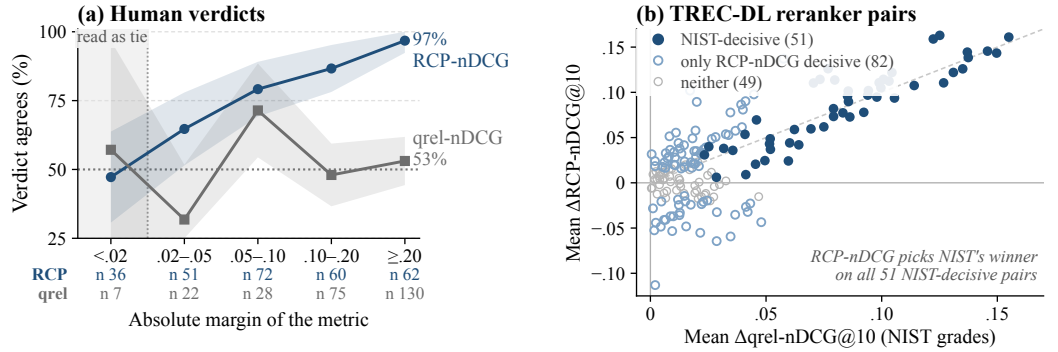


Figure 7: (a) Share of contests (two rerankers' top-five lists) whose majority verdict sides with a metric's winner, by that metric's absolute margin between the lists (95% CIs, n per band, chance 50%). Left of the dotted line: margins below 0.02, read as ties. (b) TREC-DL reranker pairs of both years (judge QWEN3.6-27B), oriented so that NIST favors the first.

about as many reranker pairs as RCP-nDCG (Figure 6a). The rubric thus supplies most of the added separation (Appendix I.3). Within a query, however, the tournament orders documents that Count-nDCG ties. Each of three blind LLM judges from three model families prefers RCP-nDCG's ideal top-10 list over Count-nDCG's on every suite, in 80.6–98.8% of queries with a preference (Figure 6b; Appendix G.7). Both components are thus needed. The criteria supply the graded signal that the qrels lack (Proposition I.4), and the tournament resolves finer differences (Proposition I.5).

4.5 Blind External Preference Validation

Finally, we test whether RCP-nDCG prefers the same rerankers as human judges.

Human verdicts. The annotators' verdicts side with RCP-nDCG@5 more often than with qrel-nDCG@5. In 185 contests, exactly one metric favors the verdict. That metric is RCP-nDCG in 72.4% of them (134 of 185; 95% CI [66.0, 78.9]%). Chance is about 53%, not 50%, because qrel-nDCG ties more often and a contest in which one metric ties can count only for the other. The share stays above chance on each suite alone (Appendix H.3) and without the contests of the zerank rerankers, which train on LLM preferences (68.1%, against 76.9% on their contests).

Metric margins. A metric's margin is its score difference between the two rerankers. In the human study, only RCP-nDCG's margins predict the annotators' verdicts (Figure 7a). A metric that tracks the annotators should match the verdict more often as its margin grows. When a metric's margin is 0.20 or more, the verdict matches its winner in 96.8% of contests for RCP-nDCG and in 53.1% for qrel-nDCG. Below a margin of 0.02, it matches RCP-nDCG's winner in only 47.2%, near chance. We therefore read such margins as ties.

NIST grades. RCP-nDCG@10 agrees with NIST on every reranker pair that NIST's grades separate (Figure 7b; Appendix H.1). We call a pair NIST-decisive when qrel-nDCG@10 with the NIST grades as linear gains separates it (paired t-test). Only 14 of the 91 pairs are NIST-decisive in 2019 and 37 in 2020. RCP-nDCG@10 favors NIST's winner on each of them with either TREC-DL judge. RCP-nDCG@10 also separates many more pairs, 60 in 2019 and 70 in 2020 with QWEN3.6-27B.

In sum, both human references favor the rerankers that RCP-nDCG favors. When exactly one metric agrees with our annotators, it is RCP-nDCG in 72.4% of contests. Only its margins predict their verdicts. It agrees with NIST on every NIST-decisive pair while separating many more pairs. RCP-nDCG thus compares rerankers better than qrel-nDCG.

5 Related Work

Item Response Theory. IRT places respondents and test items on one latent scale (Rasch, 1980; Birnbaum, 1968; Samejima, 1969; Lord, 1980). In NLP, it builds evaluation scales, compresses benchmark suites, and guides model search (Lalor et al., 2016; Maia Polo et al., 2024; Mencattini et al., 2025). These works treat models as respondents and benchmark questions as items. RCP instead treats documents as respondents and rubric criteria as items, which links the queries.

LLMs in retrieval. LLMs rerank through listwise and pairwise prompting (Sun et al., 2023; Qin et al., 2024). As judges, they remain sensitive to bias and prompting (Liu et al., 2023; Wang et al., 2024; Rahmani et al., 2025), and their use as gold labels remains debated (Thomas et al., 2024; Soboroff, 2025). Yan et al. (2024) adjust pointwise LLM labels to respect pairwise preferences. Concurrent work rescales each query's BT scores to a fixed range (Georgiev et al., 2026). RCP instead keeps the tournament order and sets each query's scale and offset from shared criteria (Appendix L).

6 Conclusion

Sparse, discrete human labels make nDCG saturate or tie close rerankers. LLM judgments could replace these labels, but relative judgments share no scale across queries, and absolute ones are too coarse. We introduced RCP and its retrieval metric, RCP-nDCG. RCP fuses relative and rubric-based LLM judgments through Item Response Theory. Against 46 annotators' grades, RCP's gains reach an AUC of 0.910, versus 0.651 for the qrels. When exactly one of RCP-nDCG and qrel-nDCG agrees with the annotators, it is RCP-nDCG in 72.4% of 185 contests (chance about 53%). With either TREC-DL judge, RCP-nDCG favors NIST's winner on every NIST-decisive pair (Section 4.5).

Each component of RCP has a separate role. The criteria define an absolute, interpretable relevance scale and carry most of RCP-nDCG's added separation of reranker pairs over qrel-nDCG (Section 4.4). The IRT merge places all queries on one scale and raises the Spearman correlation of a query's mean score with its annotators' mean grade from 0.538 to 0.795 (Section 4.2). The tournament resolves differences too fine for the rubric alone. Three blind LLM judges prefer RCP-nDCG's ideal top-10 over the rubric-only Count-nDCG's on every suite, in 80.6–98.8% of the queries with a preference (Figure 6b). We therefore suggest that retrieval benchmarks label their candidate pools once with RCP and report RCP-nDCG next to qrel-nDCG.

Limitations. RCP's gains and our annotators' grades value each document alone and ignore redundancy between documents, as on HOTPOTQA. Both measure perceived relevance, not factual correctness. The annotators' questions resemble the rubric by design. Relevance is hard to judge consistently, even for NIST assessors (Voorhees, 2000). Two annotators grade a document within one step of each other in 77.8% of cases. LLM judges may share the evaluated rerankers' preferences. Models fit on RCP signal are best tested on other labels (Appendix A).

Beyond retrieval. Arenas and LLM verifiers also pool judgments made one prompt at a time (Zheng et al., 2023; Chiang et al., 2024), and a rubric shared across prompts could calibrate them in the same way. A multidimensional IRT model could split the calibrated score $\tilde{\theta}$ into dimensions such as correctness and safety, each with its own criteria, and compare candidates by Pareto dominance. As rubric-based RL becomes central to LLM post-training, the calibrated gain could also serve as a continuous reward that rises fastest where the rubric measures most precisely (Lemma I.1).

AI USE STATEMENT

The LLM judges are the object of this study. QWEN3.5-397B, QWEN3.6-27B, and GPT-OSS-120B produce the relevance judgments that our method calibrates (Appendix C.4). Other LLMs, namely GLM-5.3-FLASH, DEEPSEEK-4.1-FLASH, KIMI K3, and GLM-5.2, serve only as blind comparison judges and auxiliary checks and never produce a gain (Appendix C.4). In this work, we used generative AI tools to implement methods, to clean and reformat datasets, and to support data analysis. These tools also helped formulate the mathematical claims of Appendix I and write their proofs. We discussed our conceptual framework, the research methodology, and the experiments with these tools to improve them. A coding agent also edited the wording of the rubric criteria during rubric development (Appendix J). We have not used generative AI tools to devise the conceptual framework, to generate synthetic data sets, to propose or refine hypotheses, or to interpret results. Translation is not applicable to this work. Additionally, we used generative AI tools to create or edit software code and figures, to draft and edit parts of the paper for readability, and to identify and validate references. No LLM annotation replaced the human annotation of the human study. We have reviewed all AI-assisted work and take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

The human study used 46 paid external contractors, who received $35 per hour and gave informed consent. The study had no formal review by an institutional review board. In the released data, one-way hashes replace the annotators' identities, and free-text comments are omitted (Appendix K). All datasets are public and used under their stated licenses or terms of use (Appendix K). The LLM judges may share biases with other LLM-based evaluation systems (Appendix A).

REPRODUCIBILITY STATEMENT

The method, the calibration fit, the evaluation protocol, and the judge budgets are specified in Appendix C. Appendix J prints the Stage A and Stage B prompts and summarizes the prompts of the comparison judges. The window schedules use a fixed seed (Appendix C). The supplementary material holds the code of both stages, the calibration, the metric, the prompts, and all judged scores, and we will release them publicly (Appendix K).

REFERENCES

Negar Arabzadeh, Alexandra Vtyurina, Xinyi Yan, and Charles L. A. Clarke. Shallow pooling for sparse labels. *Information Retrieval Journal*, 25(4):365–385, 2022. doi: 10.1007/s10791-022-09411-0.

Minjeong Ban, Jeonghwan Choi, Hyangsuk Min, Nicole Hee-Yeon Kim, Minseok Kim, Jae-Gil Lee, and Hwanjun Song. Completing missing annotation: Multi-agent debate for accurate and scalable relevant assessment for IR benchmarks. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/31f527006510a66f6accf141483efa4c-Abstract-Conference.html.

Allan Birnbaum. Some latent trait models and their use in inferring an examinee's ability. In Frederic M. Lord and Melvin R. Novick (eds.), *Statistical Theories of Mental Test Scores*, pp. 397–479. Addison-Wesley, Reading, MA, 1968.

Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. doi: 10.1093/biomet/39.3-4.324.

Chris Buckley and Ellen M. Voorhees. Retrieval evaluation with incomplete information. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 25–32. ACM, 2004. doi: 10.1145/1008992.1009000.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot Arena: An open platform for evaluating LLMs by human preference. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 8359–8388. PMLR, 2024. URL https://proceedings.mlr.press/v235/chiang24b.html.

Charles L. A. Clarke, Alexandra Vtyurina, and Mark D. Smucker. Assessing top-k preferences. *ACM Transactions on Information Systems*, 39(3):1–21, 2021. doi: 10.1145/3451161.

Cohere. Introducing Rerank 4: Cohere's most powerful reranker yet. Cohere blog, December 2025. URL https://cohere.com/blog/rerank-4. Published December 11, 2025. Model identifiers `rerank-v4.0-pro` and `rerank-v4.0-fast` as listed in the Cohere release notes, https://docs.cohere.com/changelog/rerank-v4.0.

Contextual AI. Open-sourcing reranker v2. Contextual AI blog, August 2025. URL https://contextual.ai/blog/rerank-v2. Published August 27, 2025. Models: `ctxl-rerank-v2-instruct-multilingual-{1b,2b,6b}`.

Gordon V. Cormack, Charles L. A. Clarke, and Stefan Buettcher. Reciprocal rank fusion outperforms Condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 758–759. ACM, 2009. doi: 10.1145/1571941.1572114.

Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Ellen M. Voorhees. Overview of the TREC 2019 deep learning track. In *Proceedings of the Twenty-Eighth Text REtrieval Conference (TREC 2019)*, volume 500-331 of *NIST Special Publication*. National Institute of Standards and Technology, 2020. URL https://trec.nist.gov/pubs/trec28/papers/OVERVIEW.DL.pdf. arXiv:2003.07820.

Nick Craswell, Bhaskar Mitra, Emine Yilmaz, and Daniel Campos. Overview of the TREC 2020 deep learning track. In *Proceedings of the Twenty-Ninth Text REtrieval Conference (TREC 2020)*, volume 1266 of *NIST Special Publication*. National Institute of Standards and Technology, 2021. URL https://trec.nist.gov/pubs/trec29/papers/OVERVIEW.DL.pdf. arXiv:2102.07662.

Guglielmo Faggioli, Laura Dietz, Charles L. A. Clarke, Gianluca Demartini, Matthias Hagen, Claudia Hauff, Noriko Kando, Evangelos Kanoulas, Martin Potthast, Benno Stein, and Henning Wachsmuth. Perspectives on large language models for relevance judgment. In *Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval*, pp. 39–50, Taipei, Taiwan, 2023. ACM. doi: 10.1145/3578337.3605136.

Naghmeh Farzi and Laura Dietz. Pencils down! Automatic rubric-based evaluation of retrieve/generate systems. In *Proceedings of the 2024 ACM SIGIR International Conference on Theory of Information Retrieval*, pp. 175–184. ACM, 2024. doi: 10.1145/3664190.3672511.

Naghmeh Farzi and Laura Dietz. Criteria-based LLM relevance judgments. In *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR)*, pp. 254–263. ACM, 2025. doi: 10.1145/3731120.3744591.

Nikolay Georgiev, Maria Drencheva, Kseniia Ibragimova, Ivo Petrov, Dimitar I. Dimitrov, and Martin Vechev. SABER-Math: Automated benchmark for information retrieval evaluation in mathematics. *arXiv preprint arXiv:2606.29894*, 2026. URL https://arxiv.org/abs/2606.29894. Accepted at EMNLP 2026.

Helia Hashemi, Jason Eisner, Corby Rosset, Benjamin Van Durme, and Chris Kedzie. LLM-Rubric: A multidimensional, calibrated approach to automated evaluation of natural language texts. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13806–13834, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.745.

Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4):422–446, 2002. doi: 10.1145/582415.582418.

Omar Khattab and Matei Zaharia. ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 39–48. ACM, 2020. doi: 10.1145/3397271.3401075.

Terry K. Koo and Mae Y. Li. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2):155–163, 2016. doi: 10.1016/j.jcm.2016.02.012.

Klaus Krippendorff. *Content Analysis: An Introduction to Its Methodology*. SAGE Publications, Thousand Oaks, CA, fourth edition, 2018. ISBN 978-1-5063-9566-1.

John P. Lalor, Hao Wu, and Hong Yu. Building an evaluation scale using item response theory. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 648–657, Austin, Texas, 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1062.

Sander Land and Daniel M. Bikel. Auditing LLM benchmarks with item response theory. *arXiv preprint arXiv:2605.30504*, 2026. URL https://arxiv.org/abs/2605.30504. Accepted at EMNLP 2026.

Yukyung Lee, JoongHoon Kim, Jaehee Kim, Hyowon Cho, Jaewook Kang, Pilsung Kang, and Najoung Kim. CheckEval: A reliable LLM-as-a-judge framework for evaluating text generation using checklists. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 15771–15798, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.796.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511–2522, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.153.

Yuxuan Liu, Tianchi Yang, Shaohan Huang, Zihan Zhang, Haizhen Huang, Furu Wei, Weiwei Deng, Feng Sun, and Qi Zhang. Calibrating LLM-based evaluator. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 2638–2656, Torino, Italia, 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.237/.

António Loison, Quentin Macé, Antoine Edy, Victor Xing, Tom Balough, Gabriel de Souza P. Moreira, Bo Liu, Manuel Faysse, Céline Hudelot, and Gautier Viaud. ViDoRe V3: A comprehensive evaluation of retrieval augmented generation in complex real-world scenarios. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16570–16600, San Diego, California, United States, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.755.

Kehan Long, Shasha Li, Chen Xu, Jintao Tang, and Ting Wang. Precise zero-shot pointwise ranking with LLMs through post-aggregated global context information. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2384–2394. ACM, 2025. doi: 10.1145/3726302.3730061.

Frederic M. Lord. *Applications of Item Response Theory to Practical Testing Problems*. Lawrence Erlbaum Associates, Hillsdale, NJ, 1980.

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinyBenchmarks: Evaluating LLMs with fewer examples. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 34303–34326. PMLR, 2024. URL https://proceedings.mlr.press/v235/maia-polo24a.html.

Tommaso Mencattini, Robert Adrian Minut, Donato Crisostomi, Andrea Santilli, and Emanuele Rodolà. MERGE^3: Efficient evolutionary merging on consumer-grade GPUs. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 43694–43715. PMLR, 2025. URL https://proceedings.mlr.press/v267/mencattini25a.html.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 2014–2037, Dubrovnik, Croatia, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.eacl-main.148.

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, co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016)*, volume 1773 of *CEUR Workshop Proceedings*, Barcelona, Spain, 2016. CEUR-WS.org. URL https://ceur-ws.org/Vol-1773/CoCoNIPS_2016_paper9.pdf.

Rodrigo Nogueira and Kyunghyun Cho. Passage re-ranking with BERT. *arXiv preprint arXiv:1901.04085*, 2019. URL https://arxiv.org/abs/1901.04085.

NVIDIA. Qwen3.5-397B-A17B-NVFP4. Hugging Face model card, February 2026. URL https://huggingface.co/nvidia/Qwen3.5-397B-A17B-NVFP4. Released February 17, 2026. NVFP4-quantized version of Qwen3.5-397B-A17B.

OpenAI. gpt-oss-120b & gpt-oss-20b Model Card. arXiv preprint arXiv:2508.10925, 2025. URL https://arxiv.org/abs/2508.10925.

Andrew Parry, Maik Fröbe, Harrisen Scells, Ferdinand Schlatt, Guglielmo Faggioli, Saber Zerhoudi, Sean MacAvaney, and Eugene Yang. Variations in relevance judgments and the shelf life of test collections. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 3387–3397. ACM, 2025. doi: 10.1145/3726302.3730308.

Nicholas Pipitone, Ghita Houir Alami, Advaith Avadhanam, Anton Kaminskyi, and Ashley Khoo. zELO: ELO-inspired training method for rerankers and embedding models. *arXiv preprint arXiv:2509.12541*, 2025. URL https://arxiv.org/abs/2509.12541.

Zhen Qin, Rolf Jagerman, Kai Hui, Honglei Zhuang, Junru Wu, Le Yan, Jiaming Shen, Tianqi Liu, Jialu Liu, Donald Metzler, Xuanhui Wang, and Michael Bendersky. Large language models are effective text rankers with pairwise ranking prompting. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 1504–1518, Mexico City, Mexico, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.97.

Qwen Team. Qwen3.5-397B-A17B. Hugging Face model card, February 2026a. URL https://huggingface.co/Qwen/Qwen3.5-397B-A17B.

Qwen Team. Qwen3.6-27B. Hugging Face model card, April 2026b. URL https://huggingface.co/Qwen/Qwen3.6-27B. FP8 release: https://huggingface.co/Qwen/Qwen3.6-27B-FP8.

Hossein A. Rahmani, Clemencia Siro, Mohammad Aliannejadi, Nick Craswell, Charles L. A. Clarke, Guglielmo Faggioli, Bhaskar Mitra, Paul Thomas, and Emine Yilmaz. Judging the judges: A collection of LLM-generated relevance judgements. *arXiv preprint arXiv:2502.13908*, 2025. URL https://arxiv.org/abs/2502.13908.

Georg Rasch. *Probabilistic Models for Some Intelligence and Attainment Tests*. University of Chicago Press, Chicago, expanded edition, 1980.

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009. doi: 10.1561/1500000019.

Tetsuya Sakai. Evaluating evaluation metrics based on the bootstrap. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 525–532. ACM, 2006. doi: 10.1145/1148170.1148261.

Tetsuya Sakai. Alternatives to Bpref. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 71–78. ACM, 2007. doi: 10.1145/1277741.1277756.

Fumiko Samejima. Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, 34(S1):1–97, 1969. doi: 10.1007/BF03372160. Psychometrika Monograph Supplement No. 17.

Patrick E. Shrout and Joseph L. Fleiss. Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2):420–428, 1979. doi: 10.1037/0033-2909.86.2.420.

Ian Soboroff. Don't use LLMs to make relevance judgments. *Information Retrieval Research*, 1(1): 29–46, 2025. doi: 10.54195/irrj.19625.

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 *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/7a0f8055c838df8e62329a76c7c6403d-Abstract-Conference.html.

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. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 14918–14937, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.923.

Rikiya Takehi, Ellen M. Voorhees, Tetsuya Sakai, and Ian Soboroff. LLM-assisted relevance assessments: When should we ask LLMs for help? In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 95–105. ACM, 2025. doi: 10.1145/3726302.3729916.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/65b9eea6e1cc6bb9f0cd2a47751a186f-Abstract-round2.html.

Paul Thomas, Seth Spielman, Nick Craswell, and Bhaskar Mitra. Large language models can accurately predict searcher preferences. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1930–1940. ACM, 2024. doi: 10.1145/3626772.3657707.

Ellen M. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing & Management*, 36(5):697–716, 2000. doi: 10.1016/S0306-4573(00)00010-8.

Voyage AI. rerank-2.5 and rerank-2.5-lite: instruction-following rerankers. Voyage AI blog, August 2025. URL https://blog.voyageai.com/2025/08/11/rerank-2-5/. Published August 11, 2025.

Feng Wang, Yuqing Li, and Han Xiao. jina-reranker-v3: Last but not late interaction for listwise document reranking. *arXiv preprint arXiv:2509.25085*, 2025. URL https://arxiv.org/abs/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 *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9440–9450, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.511.

Jacob Whitehill, Paul Ruvolo, Tingfan Wu, Jacob Bergsma, and Javier Movellan. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Advances in Neural Information Processing Systems*, volume 22, pp. 2035–2043, 2009. URL https://proceedings.neurips.cc/paper_files/paper/2009/hash/f899139df5e1059396431415e770c6dd-Abstract.html.

Le Yan, Zhen Qin, Honglei Zhuang, Rolf Jagerman, Xuanhui Wang, Michael Bendersky, and Harrie Oosterhuis. Consolidating ranking and relevance predictions of large language models through post-processing. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 410–423, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.25.

Seonghyeon Ye, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, Seungone Kim, Yongrae Jo, James Thorne, Juho Kim, and Minjoon Seo. FLASK: Fine-grained language model evaluation based on alignment skill sets. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/f41b4a6b202adcd8e150a9d4f124d8f6-Abstract-Conference.html.

Wendy M. Yen. Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, 8(2):125–145, 1984. doi: 10.1177/014662168400800201.

Soyoung Yoon, Gyuwan Kim, Gyu-Hwung Cho, and Seung-won Hwang. AcuRank: Uncertainty-aware adaptive computation for listwise reranking. In *Advances in Neural Information Processing Systems*, volume 38, pp. 29944–29973, 2025. doi: 10.52202/085713-0890.

Zeta Alpha. NanoBEIR. https://huggingface.co/collections/zeta-alpha-ai/nanobeir-66e1a0af21dfd93e620cd9f6, 2024.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025. URL https://arxiv.org/abs/2506.05176.

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 (Datasets and Benchmarks Track)*, volume 36, pp. 46595–46623, 2023. doi: 10.52202/075280-2020.

Honglei Zhuang, Zhen Qin, Kai Hui, Junru Wu, Le Yan, Xuanhui Wang, and Michael Bendersky. Beyond yes and no: Improving zero-shot LLM rankers via scoring fine-grained relevance labels. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pp. 358–370, Mexico City, Mexico, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-short.31.

Justin Zobel. How reliable are the results of large-scale information retrieval experiments? In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 307–314. ACM, 1998. doi: 10.1145/290941.291014.

Appendices

A Limitations **17**

B Definitions **17**

C Method Details **18**

D Human Study Protocol **22**

E Diagnostics and Calibration **24**

F Agreement with Ground Truth and Human Labels **28**

G Ranking Results **30**

H External Preference Validation **37**

I Theory **42**

J Prompts **44**

K Assets and Release **48**

L Extended Related Work **49**

Reader's guide. The table maps each main-text section to the appendix sections that hold its details and robustness checks. After the limitations, definitions, and method details, the appendix follows the order of Section 4.

Main text	Claim or component	Appendix
Section 3	Protocols, calibration fit, judges, budgets, and cost; prompts; proofs	Appendix C, Appendix J, Appendix I
Section 4	Human study: design, what annotators saw, annotators, reliability	Appendix D
Section 4.1	Criterion parameters, per-query τ_j and α_j, transfer, label-free checks	Appendix E.1, Appendix E.2
Section 4.2	Gain against human grades; calibrated scores against simpler maps; NIST placement; NIST positives that RCP scores lowest	Appendix E.3, Appendix E.4, Appendix H.2
Section 4.3	Study 1 (qrel sets against RCP sets); gain against human labels	Appendix F.2, Appendix F.1
Section 4.4	Failure modes, sensitivity and scores per dataset, second judges, zerank, robustness, Count-nDCG	Appendix G.1, Appendix G.2, Appendix G.3, Appendix G.4, Appendix G.5, Appendix G.6, Appendix G.7
Section 4.5	Human verdicts and their controls, NIST-decisive pairs	Appendix H.3, Appendix H.1
Section 5	Extended related work	Appendix L
Section 6	Limitations; HOTPOTQA	Appendix A, Appendix F.3
Appendix only	Study 2 (reranker lists judged by LLMs)	Appendix H.4

A LIMITATIONS

What the metric measures. Since RCP's gains and our annotators' grades value each document on its own, they do not penalize redundancy. On HOTPOTQA, whose questions need a primary and a secondary fact, a redundant document on the primary fact can outrank the document with the secondary one (Appendix F.3). Pointwise gains also cannot credit a document that becomes relevant only once another document is known. The judge and the annotators both measure perceived relevance, not factual correctness (Appendix D.2). When the RCP-nDCG@5 scores of two lists differ by less than 0.02, annotators side with RCP-nDCG only at chance. We therefore read such differences as ties (Appendix H.3).

The evidence. Our document-level and contest-level human results share annotations. The NIST grades share none with either. Relevance is hard to judge consistently, even for NIST assessors (Voorhees, 2000). Two annotators of the same document grade it within one step of each other in 77.8% of cases. Gwet's AC2 with quadratic weights, which credits near misses on the ordinal scale, is 0.630, and Krippendorff's α with interval distances is 0.571. Both are 0 at chance and 1 at perfect agreement. We therefore label documents by the mean of at least three grades. The annotators' questions resemble the rubric by design (Appendix D.2). The external LLM judges of Appendices G.7 and H.4 may share systematic preferences with our judge (Soboroff, 2025). The NIST check replaces our primary judge with the newer and smaller QWEN3.6-27B and adds GPT-OSS-120B. With either judge, RCP-nDCG favors NIST's winner on every NIST-decisive pair (Appendix H.1). On zerank-2, which learns from LLM preferences, unjudged passages explain only part of its lower rank under the NIST grades (Section 4.4).

Scope and cost. We refined the criteria on 15 NanoBEIR queries with 20 documents each, about 0.3% of its pooled documents (Appendix J). All other data are held out. The criterion parameters also transfer to datasets left out of the fit (Appendix E.1). Scores are comparable only within one calibration fit. The fit also ignores the uncertainty of the Stage A scores. Documents outside the judged pool need new judgments and a refit (Appendix C.4). We validate RCP as an evaluation method. Models trained on its feedback are best tested on other labels, since RCP-nDCG would reward their own training signal. This concern may be smaller than it seems. Rerankers are far smaller than the LLM judge and spend far less computation per document. Such rerankers are thus unlikely to reach the judge's quality. Future work can test whether RCP-nDCG remains informative for models trained on its feedback.

B DEFINITIONS

B.1 METRICS AND TIES

This appendix defines the metrics, statistics, and intervals that the other appendices use. All four compared metrics are nDCG (Equation (1)) and differ only in the gain. nDCG on the qrels (qrel-nDCG) uses the qrel grades as gains. Count-nDCG uses the pass share of Equation (5). Raw-BT-nDCG applies the gain g of Equation (4) to the uncalibrated BT scores, centered at zero within each query. RCP-nDCG uses the calibrated gain of Equation (4). Reranker scores can tie. On ViDoRe v3, each document in a tied group receives the group's mean gain. On NanoBEIR and BRIGHT, we break tied scores by document identifier (the trec_eval convention). Giving tied documents their group's mean gain changes the sensitivity of qrel-nDCG and RCP-nDCG (Appendix B.2) by at most 0.4 points. On TREC-DL, tied scores keep their order in the judge pool, which starts with the fused first-stage ranking. On NanoBEIR and BRIGHT, the failure-mode shares instead keep the first-stage order of tied documents (Appendix G.1).

B.2 STATISTICS

A metric separates two rerankers on a dataset when a paired t-test on their per-query scores rejects equal means at the 5% level. Sensitivity is the share of the 91 reranker pairs of a dataset that a metric separates, averaged over the datasets of a suite. If no two rerankers differed, about 5% of pairs would separate by chance. The AUC is the share of pooled document pairs, one rated Useful or better and one not, that a score orders correctly, with ties counting half (0.5 is chance, 1 is perfect).

Within-query concordance is the share of same-query pairs with different grades that the score orders like the grades (0.5 is chance).

The expected calibration error (ECE) is the size-weighted mean gap between predicted probabilities and observed frequencies in bins of the prediction (0 is perfect). The Brier skill score is the share of a constant prediction's squared error that the pass probabilities remove (1 is perfect, 0 is no better than the constant).

ICC(1,1) is the expected correlation between two single grades of one document (Shrout & Fleiss, 1979; Koo & Li, 2016). ICC(1,k) is the expected correlation between the mean grades of two independent panels of k annotators. The rank-based split-half reliability correlates the mean grades of two random half-panels and steps the correlation up to the full panel with the Spearman–Brown formula. Krippendorff's α with the interval metric is one minus the ratio of observed to chance-expected squared grade differences (0 is chance, 1 is perfect agreement) (Krippendorff, 2018). On our grades it is 0.571, nearly equal to ICC(1,1) at 0.572. Fleiss' κ compares the observed agreement of annotators with the agreement that their answer rates produce by chance (0 is chance, 1 is perfect agreement). Gwet's AC1 also corrects for chance but depends less on how common each answer is.

A contest's verdict is the list that more annotators pick. A contest is decided when its verdict names a list that exactly one of the two metrics also names. The main measure is the share of decided contests whose verdict sides with RCP-nDCG. The chance level of the main measure is 52.7%, not one half, because qrel-nDCG ties more often (Appendix H.3). A contest in which one metric ties can count only for the other. A metric's margin is its score difference between list A and list B. A metric ties when the absolute value of its margin is below 0.005. The margin AUC of a metric is the probability that a random annotator pick of list A has a larger A-minus-B margin than a random pick of list B (0.5 means no information).

B.3 Intervals

Unless the text or a caption says otherwise, intervals are 95% query-clustered percentile bootstraps with 2,000 resamples and a fixed seed, stratified by suite, corpus, or year. Rates over human contests, such as the main measure, instead use an analytic query-clustered interval. The margin AUC uses a delete-one-query jackknife. All methods treat the query as the unit, because documents and contests of one query share its pool.

C Method Details

C.1 Protocol, pools, and rerankers

This appendix details the candidate pools, the judging schedules, the calibration fit, the judges and their cost, and the setup of each suite. Table 2 collects the settings of the full protocol, the Stage A and Stage B schedules that NanoBEIR and BRIGHT use (Appendix C.2).

Table 2: Settings of the full protocol, which NanoBEIR and BRIGHT use under both judges. Appendix C.7 describes the reduced protocol of TREC-DL.

Setting	Value
Pool size K (top of the fused first-stage ranking)	150
First-stage fusion	RRF over three retrievers, $\kappa = 60$
Stage A: $w = 10$, score range $[-5, +5]$, coverage windows per query (random, then stratified, each reversed)	80
Stage A: adaptive calls per query; discount of covered boundaries after each pick	56 (7 batches of 8 windows); 0.3
Stage A: BT fit	ridge 10^{-4}, L-BFGS, at most 250 iterations
Stage B: windows per query, each of $w = 10$ documents on $C = 5$ criteria, no reversed copies	100
Calibration: hard constraints	$\sum_c \gamma_c = C$, $\mathrm{mean}_c\, \beta_c = 0$
Calibration: priors on τ_j and α_j	$\mathcal{N}(1, 1)$ and $\mathcal{N}(0, 2^2)$
Calibration: ridge on criterion parameters; fit	10^{-4}; L-BFGS, at most 500 iterations
Temperature, schedule seed, metric cutoff	1.0, 42, 10

Table 3: The 14 rerankers. Families, identifier prefix, and license of the public weights: Contextual AI (Contextual AI, 2025) (`ContextualAI/`, CC BY-NC-SA 4.0), Jina AI (Wang et al., 2025) (`jinaai/`, CC BY-NC 4.0), Qwen (Zhang et al., 2025) (`Qwen/`, Apache 2.0), Cohere (Cohere, 2025) (API only), Voyage AI (Voyage AI, 2025) (API only), ZeroEntropy (Pipitone et al., 2025) (`zeroentropy/`, Apache 2.0).

Name	Public identifier	Params
Contextual v2 1B	`ctxl-rerank-v2-instruct-multilingual-1b`	1.3
Contextual v2 2B	`ctxl-rerank-v2-instruct-multilingual-2b`	2.6
Contextual v2 6B	`ctxl-rerank-v2-instruct-multilingual-6b`	6.8
Jina Reranker v3	`jina-reranker-v3`	0.6
Qwen3-Reranker-0.6B	`Qwen3-Reranker-0.6B`	0.6
Qwen3-Reranker-4B	`Qwen3-Reranker-4B`	4.0
Qwen3-Reranker-8B	`Qwen3-Reranker-8B`	8.2
Rerank 4 Fast	`rerank-v4.0-fast`	–
Rerank 4 Pro	`rerank-v4.0-pro`	–
Voyage 2.5 Lite	`rerank-2.5-lite`	–
Voyage 2.5	`rerank-2.5`	–
zerank-1-small	`zerank-1-small`	1.7
zerank-1	`zerank-1`	4.0
zerank-2	`zerank-2`	4.0

Each query's candidate pool holds the top $K = 150$ documents of a reciprocal rank fusion (RRF) of three first-stage retrievers. The three retrievers are the dense Cohere Embed v4, BM25 (Robertson & Zaragoza, 2009) as implemented in `bm25s`, and the dense Octen-Embedding-8B. Fusion scores each document d by $\sum_m 1/\big(\kappa + \mathrm{rank}_m(d)\big)$ over the three rankings m (Cormack et al., 2009). On NanoBEIR, BRIGHT, and ViDoRe v3, documents that the qrels mark relevant but the retrievers miss replace the lowest-ranked pool documents. The pool size K thus stays fixed. On TREC-DL, each pool also joins every passage that NIST assessors judged for the query. These pools are therefore larger (Appendix C.7). The pools of a few NanoBEIR and BRIGHT queries differ from this construction (Appendix C.5).

We evaluate the 14 rerankers of Table 3. The human study compares pairs drawn from 10 of them (Appendix D). The three ZeroEntropy rerankers are trained on preferences from an ensemble of LLMs (Pipitone et al., 2025). A metric built on an LLM judge could favor them. We therefore also report the human results without them.

C.2 Stage A and Stage B schedules

Stage A schedules its windows in three phases, first random, then stratified, then adaptive. Random windows give every document broad coverage. Stratified windows then group documents of similar estimated quality. Both phases show each window a second time in reverse order.

The adaptive phase adds 56 calls per query in 7 batches of 8 windows and refits the BT model before each batch. Let $p_r = \sigma\big(\hat{\theta}_{(r)} - \hat{\theta}_{(r+1)}\big)$ be the predicted preference between the documents at ranks r and $r+1$ of the current order. The phase values the boundary between them by

$$v_r = p_r\,(1 - p_r) \cdot \frac{1}{\log_2(r+1)} \cdot \frac{1}{1 + n_r}, \tag{6}$$

where n_r counts the windows that have already held both documents. The first factor peaks when the order of the two documents is a coin flip. The second factor is the rank discount of nDCG. The third favors pairs that the judge has rarely compared. A batch greedily selects the 8 windows of w consecutive documents with the largest sum of boundary values. After each pick, the values of the boundaries it covers shrink by the factor of Table 2. Adaptive windows are shown once, in the current estimated order.

The w scores of a window fix only $w - 1$ independent differences. Each of its $\binom{w}{2}$ pairs therefore has weight $2/w$. Each BT fit minimizes the weighted mean cross-entropy between soft and predicted preferences plus the ridge penalty of Table 2. If a response's scores cannot be parsed, its ranking enters the fit as hard preferences with weight 1.

Stage B also shows w documents per call and asks about each on its own (Section 3.2). Under the full protocol and on ViDoRe v3, each query receives 100 windows, whose first 50 are balanced random windows. Their sampling is inspired by balanced incomplete block designs. Each window takes the documents shown least often so far, with random tie-breaks. All documents thus appear about equally often and meet many different partners. A Rasch model (Rasch, 1980), the one-parameter case of IRT, then gives each document a preliminary estimate from these responses. The remaining windows group documents with similar estimates. The calibration treats each of a document's responses as one observation (Appendix C.3). We call the three Stage A phases together with this Stage B schedule the full protocol, which NanoBEIR and BRIGHT use (Table 2). ViDoRe v3 uses its Stage B schedule (Appendix C.6).

C.3 Calibration fit

Within a query, the ability is a positive affine function of the BT score. This affine map keeps the query's order and the model linear on the logit scale (Equation (2)). We set $\tau_j = \mathrm{softplus}(t_j)$ with an unconstrained t_j, so that no negative scale can reverse the tournament order. The discriminations are $\gamma_c = C\, e^{u_c} / \sum_{c'} e^{u_{c'}}$ and the difficulties are $\beta_c = v_c - \mathrm{mean}_{c'}\, v_{c'}$. Let N be the number of Stage B responses, each of which answers all C criteria. The fit minimizes

$$\frac{1}{NC}\left[\sum_{n=1}^{N}\sum_{c=1}^{C}\ell_{nc} + \sum_{j}\left(\frac{(\tau_j - 1)^2}{2\sigma_\tau^2} + \frac{\alpha_j^2}{2\sigma_\alpha^2}\right)\right] + \frac{\lambda}{2}\sum_{c}\left(u_c^2 + v_c^2\right), \tag{7}$$

where ℓ_{nc} is the binary cross-entropy of response n on criterion c under Equation (2) and the hyperparameters take the values of Table 2. Up to the factor $1/(NC)$, the objective is a negative log posterior under the priors of Table 2, with the ridge acting as a Gaussian prior on u_c and v_c. The fit is therefore the maximum a posteriori (MAP) estimate. The priors keep the estimates finite when all documents of a query fail every criterion or all pass every criterion.

We hold $\hat{\theta}^{\mathrm{BT}}$ fixed from Stage A and do not propagate their uncertainty (Appendix A). The fit starts from $\tau_j = 1$, $\alpha_j = 0$, and equal criterion parameters. On ViDoRe v3, the fit covers all six language versions of every question. On TREC-DL, the main fit pools both years with further MS MARCO queries without NIST grades (Appendix C.7).

C.4 Judges, budgets, and cost

Judges and serving. We serve the open-weight judges with SGLang, each on one node of eight GPUs. The primary judge, Qwen3.5-397B (Qwen Team, 2026a), judges NanoBEIR, BRIGHT, and ViDoRe v3 in its 4-bit NVFP4 release (NVIDIA, 2026) at public revision `0368c1b` on eight B200 GPUs. The second judge, gpt-oss-120b (OpenAI, 2025), runs in its released precision (MXFP4 mixture-of-experts weights) on H100 GPUs and repeats both stages on NanoBEIR, BRIGHT, and TREC-DL. On TREC-DL, the smaller dense Qwen3.6-27B (Qwen Team, 2026b) judges in its 8-bit (FP8) release under the reduced protocol of Appendix C.7. No other LLM produces a gain. The other LLMs serve as blind comparison judges (Appendices H.4, H.5, and J.3) and as a TREC-DL arbiter (Appendix G.4).

Budgets and cost. Table 4 lists how often the judge sees each document. A reversed copy counts as a separate call. A placement is one appearance of a document in a Stage B window.

Each pool is judged once and then scores any reranking of it without further judge calls. A document outside the judged pool needs new windows and a refit of the 2PL model, which can shift earlier calibrated scores. The cost of RCP is thus paid per pool, as for human qrels, not per reranker.

We price the logged tokens at public API rates. The TREC-DL run of Qwen3.6-27B costs \$360.91 in total and \$3.72 per query. This total includes imputed reasoning tokens for the 2020 queries, which make up 34.7% of it. We also replay the logged judgments at two reduced budgets, which keep 50.2% and 35.0% of the document judgments. These replays cost \$1.87 and \$1.33 per query. Each replay draws five random subsamples and refits both stages on each. The replays and the full leaderboard they are compared with fit the calibration on the TREC-DL queries alone. The replays score the 18 TREC-DL systems, namely the 14 rerankers, the three retrievers, and their fusion. We report TREC-DL 2019, whose changes exceed those of 2020. With half of the judgments, the largest

Table 4: Judge budgets per suite. Calls and Stage B windows are counted per query. Windows and placements per document are means over pool documents. GPT-OSS-120B uses the settings of QWEN3.5-397B on NanoBEIR and BRIGHT.

		Stage A			Stage B	
Suite	Judge	Calls	Reversed	Windows/doc	Windows	Placements/doc
NanoBEIR	QWEN3.5-397B	216	yes	14.4	100	6.7
BRIGHT	QWEN3.5-397B	216	yes	14.4[a]	100	6.7[a]
ViDoRe v3	QWEN3.5-397B	–[b]	–[b]	–[b]	100	6.67
TREC-DL 2019	QWEN3.6-27B	325	no	11.27	113[c]	3.92
TREC-DL 2020	QWEN3.6-27B	375	no	13.08	113[c]	3.94

[a]Computed from the configuration. The released window logs give the same value. [b]Not recorded in the stored run artifacts. [c]Mean over the queries of both years, whose Stage B windows are pooled.

change in a system's mean RCP-nDCG@10 is 0.009, averaged over subsamples. The Kendall rank correlation between the reduced and the full leaderboard is 0.987, where 1 means the same order. With 35.0% of the judgments, these values are 0.026 and 0.969. Neither budget reverses a system pair that the full budget separates. This check therefore compares RCP-nDCG only with itself, not with NIST.

C.5 NANOBEIR AND BRIGHT

All RCP scores on NanoBEIR and BRIGHT come from QWEN3.5-397B, except where we name GPT-OSS-120B. Appendix E.5 compares gain functions on these suites. Appendix B.1 describes how we break tied reranker scores, including in the failure-mode shares (Appendix G.1).

NanoBEIR. NanoBEIR (Zeta Alpha, 2024; Muennighoff et al., 2023) holds 649 queries in 13 datasets. Four NFCorpus queries have more positive qrels than K. Their reranker pools hold 196 to 463 documents, of which the judge saw 150, and some documents outside the judged pool reach the rerankers' top 10. On 7 NanoSCIDOCS queries, one positive document has empty text and is missing from the pool. qrel-nDCG still counts it in its ideal ranking. RCP-nDCG gives gain zero to a document without a calibrated score, and its ideal ranking draws on the judged pool (Equation (4)). On 20 of the 50 NanoArguAna queries, the pool holds the query argument itself, as in MTEB's evaluators. The qrels never mark this document relevant. Dropping these documents does not change the RCP-nDCG@10 leaderboard.

BRIGHT. BRIGHT has 1,384 queries in our reranker runs and 1,383 in the calibration fit. The two query sets differ only in the two TheoremQA tasks. All metrics use the 1,366 queries in both sets, except qrel-nDCG's sensitivity, which uses all TheoremQA-Theorems queries as in Figure 6.

As in BRIGHT's official evaluation (Su et al., 2025), we remove each query's listed excluded documents from every ranking we score. The removal affects 446 of the 1,366 evaluated queries. None of the removed documents is qrel-relevant. The judge's pools still contained the excluded documents, 7.1% of the Stage B observations of each fit. We keep the fitted parameters and residual checks (Tables 7 and 8). The human study is unaffected (Appendix D.1). A few follow-up analyses still use rankings that contain the excluded documents.

C.6 VIDORE V3

ViDoRe v3 (Loison et al., 2026) grades retrieval from pages of enterprise, governmental, and educational documents, which every system reads as the benchmark's OCR text, not as an image. The benchmark has eight corpora, five English and three French. Each question exists in six language versions with one shared set of relevance labels. We ran both stages of RCP on all 14,514 question rows and fitted one calibration over them. The unit of analysis is the native question (2,419 questions, Appendix E.4). Each question row has its own candidate pool of 150 pages, built as on NanoBEIR (Appendix C.1). After a pre-filter by a vision-language model, ViDoRe's human assessors graded the remaining pages on a three-point scale (Loison et al., 2026). Relevant pages thus have grade 1 or 2. On ViDoRe v3, qrel-nDCG uses these grades as linear gains. QWEN3.5-397B judges ViDoRe v3,

with 6.67 Stage B placements per page on average (Table 4). Pages with tied reranker scores receive their group's mean gain. The human study drew its contests from the English corpora (Appendix D).

Per corpus, qrel-nDCG separates 66.3% of the 91 reranker pairs on average, and RCP-nDCG separates 88.9%. The Kendall correlation between the two leaderboards, where 1 means identical orders, is 0.714. qrel-nDCG@10 saturates, floors, or compresses on 23.1% of questions (Appendix G.1).

C.7 TREC-DL 2019 AND 2020

The TREC-DL tracks of 2019 and 2020 give dense graded human labels that share no annotations with our human study. NIST assessors graded MS MARCO passages from 0 to 3 under their own guidelines, years before our method existed (Nguyen et al., 2016; Craswell et al., 2020; 2021).

Queries and pools. The two tracks have 43 and 54 queries with NIST grades, over 9,260 and 11,386 graded passages. A passage outside this judgment pool is unjudged and counts as irrelevant. Each pool joins the fused top 150 passages (Appendix C.1) with every graded passage, so that every graded passage receives an RCP gain. The mean pool holds 287.4 passages per query.

Panels and metrics. The 14 rerankers rank each query's candidates in two panels. The pool panel reorders the full pool. The top-150 panel reorders only the 150 fused passages. Leaderboards of 18 systems add the three retrievers and their fusion. We write qrel-nDCG@10 for nDCG@10 on the NIST grades with linear gain and an ideal ranking over every graded passage, as in trec_eval. The exponential gain $2^{\text{grade}} - 1$ is a reported variant. RCP-nDCG@10 and Count-nDCG@10 draw their ideal rankings from the candidate pool. We test each year separately. For each of the 91 reranker pairs, a paired t-test compares per-query scores. A metric separates the pair when the two-sided test rejects equal means at the 5% level. A pair is NIST-decisive when qrel-nDCG@10 on the NIST grades separates it. A metric agrees with NIST on such a pair when its mean difference has the same sign as NIST's.

Judge and reduced protocol. The main TREC-DL scores of RCP and Count-nDCG use the five criteria of Table 1. The reduced protocol skips the reversed Stage A copies of the full protocol and adds fewer adaptive windows per passage. A passage appears in 12.27 Stage A windows on average, against 14.4 on NanoBEIR. Stage B places each passage in 3.93 windows (Table 4).

Calibration fits. The main calibration CAL-A fits the 2PL model once over 597 queries. The fit covers the queries of both years and 500 MS MARCO development queries without NIST grades. CAL-B fits the same Stage A scores and a repeated Stage B run on the TREC-DL queries alone. CAL-C fits the judgments of the second judge (Appendix G.4). In all three fits, C3 is easier to pass than C2. C5 is far more difficult than any other criterion.

D HUMAN STUDY PROTOCOL

This appendix describes the design and sample of the human study, what annotators saw, who the annotators were, and how reliably they graded. The study has two layers. The label layer compares document grades with qrels and RCP scores (Appendix F.1). The verdict layer asks which of two lists annotators prefer (Appendix H.3). Both layers rest on the same annotations and are therefore not independent evidence.

D.1 DESIGN AND SAMPLE

The study asks whether annotators side with RCP-nDCG more often than with qrel-nDCG where exactly one of the two metrics names the annotators' choice. A contest is one query with two rerankers. Contests of the same query are correlated. Every interval therefore treats a query as one cluster (Appendix B.3).

The study covers 12 of the 13 NanoBEIR datasets, 7 of the 12 BRIGHT tasks, and 5 English-language corpora of ViDoRe v3. We exclude NanoSCIDOCS, whose qrels mark the papers that the query paper cites. We also exclude BRIGHT's five coding and theorem tasks (LeetCode, Pony, AoPS, and

both TheoremQA tasks), where relevance rests on a shared line of reasoning. The suites contribute 97, 95, and 100 distinct queries. The contests draw on 10 rerankers in 42 distinct pairs (Table 5).

We oversample disagreements, contests on which the two metrics name opposite winners. We split disagreements into decisive and marginal strata by the size of the two metrics' margins, their score differences between the two lists. The draw compares the metrics at depth 10, the usual evaluation depth, on the full rankings. Every verdict analysis compares them at depth 5 on the displayed documents. Annotators grade only each list's top five, since the union of two top-ten lists, up to 20 documents, is too long to grade with care. The sample also includes concordant contests, on which both metrics name the same winner. We targeted 100 contests per suite. The final sample holds 311 contests, 221 drawn as disagreements and 90 as concordant contests. These contests cover 292 distinct queries, because some queries enter more than one contest with different reranker pairs. Contests with a zerank reranker make up 52.1% of the sample. We also report results on the 149 contests without zerank. The first valid review was completed on 2026-07-20 and the last on 2026-08-25.

Table 5: Composition of the human study. A review is one annotator's grading of one contest. Since annotators worked across suites, their counts do not add up.

	NanoBEIR	BRIGHT	ViDoRe v3	All
Datasets or corpora	12	7	5	24
Queries	97	95	100	292
Contests	108	103	100	311
Reviews	324	309	300	933
Annotators	33	39	35	46
Grades	2,322	2,580	2,178	7,080
Distinct graded documents	723	819	726	2,268

D.2 What annotators saw

Annotators saw the query, a one-line task description of what counts as relevant, and the union of both rerankers' top five documents in random order. The annotators never saw the rerankers, the two lists, the qrels, or any metric score. For each document they answered the four questions of Table 6. Written guidelines asked them to grade only the shown text at face value and to choose Either between adjacent grades where it was offered. The questions echo the wording of four Stage B criteria (Table 1), with Topical matching C1, Useful C2, Answers C4, and Complete C5. The match lies in the wording only. We did not measure whether the questions and the criteria are equivalent.

Table 6: The annotators' four questions, in the order shown, with the grade that the most stringent yes gives and the gain of that grade in the verdicts. Answers and Complete also accept Either, which gives the half grade in parentheses and a linearly interpolated gain when it directly follows the last yes.

Name	Question	Answers	Grade	Gain
–	(every answer is no)	–	0	0
Topical	*Is this document clearly about the query's specific topic or need, not just the same broad area?*	yes, no	1	0.25
Useful	*Can you extract at least one concrete fact, step, or example that specifically addresses this query?*	yes, no	2	0.45
Answers	*Does it directly answer the query's main question or need?*	yes, Either, no	3 (2.5)	0.70 (0.575)
Complete	*Does it fully resolve the query's need on its own, with nothing important left to add?*	yes, Either, no	4 (3.5)	1 (0.85)

Annotators answer every question for every document, whatever their earlier answers. The most stringent yes therefore sets the grade. A fixed gain map (Table 6) turns grades into gains. Each annotator's grades give both lists an nDCG@5 under one shared ideal ranking. An annotator's pick is the list with the higher nDCG@5. A gap below 0.005 is a tie. A Copeland score over the annotator's rankings of equally graded documents can break such a tie. The score counts wins plus half the ties in these pairwise judgments, weighted by each document's effect on the score difference. The contest

verdict counts picks of list A against picks of list B and ignores ties (Appendix B.2). Of the 311 contests, 139 have verdict A, 150 verdict B, and 22 a tie. Appendix H.3 recomputes the verdicts under other gain maps.

A metric's margin is its nDCG@5 difference between the two displayed lists, with the ideal DCG@5 taken over the query's whole candidate pool. For RCP-nDCG, the ideal falls back to the displayed documents in 27 contests. Under the same rule, a metric names a winner only when its margin is not a tie. A contest is decided when it has a verdict and exactly one metric names the verdict's list. The main measure is the share of decided contests whose verdict sides with RCP-nDCG. In 14 contests, no document with a positive qrel appears in either top five. Table 21 also reports the main measure without these contests. On some contests one displayed document is an attention check, an off-topic document without a positive qrel and with a near-zero RCP gain. A grade above 0 on it voids that annotator's review. Both metrics call it irrelevant. Voiding a review therefore cannot favor either metric.

D.3 Annotators and terms

The annotators were 46 external contractors, each with at least a BSc in the subject of the documents they graded. They were not employees of the authors' organization or professional relevance assessors. They were paid $35 per hour and gave informed consent. The study had no review by an institutional review board. A separate team ran recruitment, quality control, and all contact with the annotators behind an information barrier. The authors wrote the questions and the guidelines. Three annotators graded each contest, 933 reviews in all. The separate team could exclude an annotator only on grounds that involve neither metric, such as failed attention checks or low agreement with other annotators. The team excluded one annotator, whose reviews we do not use and who is not among the 46 counted above. Re-admitting this annotator's 14 reviews moves the main measure from 72.4% to 72.3%. The results on disagreements alone and on disagreements with both margins at least 0.02 are unchanged (Appendix H.3).

D.4 Reliability

We measure how consistently annotators grade, first for single grades and then for the panel means and verdicts that our analyses use. Krippendorff's α for one annotator's grade on the 0 to 4 scale, where 0 is chance and 1 is perfect agreement, is 0.571 (95% CI [0.535, 0.604]). This value is below the level that Krippendorff (2018) treats as reliable. The analyses instead rest on panel means and majority verdicts. For the panel means of k annotators, the reliability ICC(1,k) (Appendix B.2) is 0.802 (95% CI [0.778, 0.823]), within the band that Koo & Li (2016) call good. Pairs of annotators of one contest agree on a pick in 72.1% of cases, where one half is chance. Fleiss' κ corrects this share for chance agreement, with 0 at chance and 1 for perfect agreement. On the picks, κ is 0.421 (95% CI [0.335, 0.505]). For a rare answer, κ can be low even when raw agreement is high. Among yes or no answers, annotators say yes to Complete in only 11.8% of cases. On these answers, κ is 0.411. Gwet's AC1, which depends less on prevalence, is 0.846 on the same answers. With any one annotator left out, α stays between 0.554 and 0.581 and verdict agreement between 70.2% and 73.8%.

E Diagnostics and Calibration

This section details the 2PL fits behind Sections 4.1 and 4.2: their parameters and transfer, label-free diagnostics, calibration against human and NIST grades, and the shape of the gain.

E.1 Fitted parameters and transfer across datasets

This subsection supports the diagnostic claims of Section 4.1. Table 7 lists the criterion parameters β_c and γ_c of the four fits on NanoBEIR and BRIGHT, one per suite and judge, shared across all queries (Section 3.3).

The per-query scale τ_j and offset α_j calibrate each query's BT scores (Equation (3)). Figure 8 shows the raw and calibrated pool scores. As BT scores are centered within each query, the offset equals the query's mean calibrated pool score. In every fit, this mean lies well below C1's difficulty (Table 7).

Table 7: Parameters of the four fits. (a) Difficulty β_c and discrimination γ_c per criterion. (b) The 5th and 95th percentiles of τ_j, their ratio, and the mean offset α_j.

(a) Criterion parameters

	NanoBEIR				BRIGHT			
	QWEN3.5-397B		GPT-OSS-120B		QWEN3.5-397B		GPT-OSS-120B	
	β_c	γ_c	β_c	γ_c	β_c	γ_c	β_c	γ_c
C1	−2.57	1.17	−2.14	1.25	−3.22	1.11	−2.68	1.15
C2	−2.00	1.05	−1.46	1.15	−1.83	1.03	−1.63	1.13
C3	−0.25	0.94	−0.25	0.84	0.70	0.73	0.31	0.73
C4	0.71	1.16	0.91	0.97	1.85	1.27	1.50	1.13
C5	4.12	0.69	2.95	0.79	2.50	0.86	2.50	0.86

(b) Per-query parameters

Suite	Judge	τ_j 5%	τ_j 95%	95%/5%	mean α_j
NanoBEIR	QWEN3.5-397B	0.86	2.72	3.2	−5.64
NanoBEIR	GPT-OSS-120B	0.95	3.18	3.3	−5.94
BRIGHT	QWEN3.5-397B	0.88	2.66	3.0	−5.22
BRIGHT	GPT-OSS-120B	1.04	3.17	3.1	−5.58

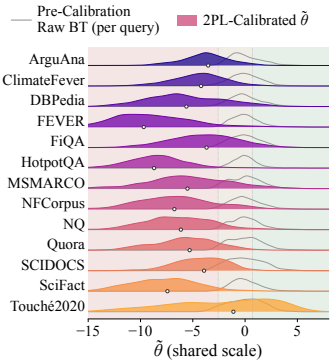


Figure 8: Per NanoBEIR dataset, pool documents' raw BT scores, centered per query (grey), and their calibrated scores (color; dot: median α_j).

To test whether the criterion parameters transfer across datasets, we fit them on all datasets but one, freeze them, and fit only τ_j and α_j on the held-out queries. We score the fit by the binary cross-entropy (BCE) of the held-out Stage B responses, where lower is better. A baseline predicts each response by its criterion's pass rate on the training datasets. On NanoBEIR with QWEN3.5-397B, the mean held-out BCE drops by 60.6% against the baseline. With GPT-OSS-120B, the reduction on NanoBEIR is 64.1%. On BRIGHT, it is 46.5% with QWEN3.5-397B and 49.6% with GPT-OSS-120B. A variant with every discrimination set to 1 and the difficulties kept (a Rasch form) raises the mean held-out BCE relative to the 2PL. The increase is 1.1–1.3% on NanoBEIR and 2.0–2.4% on BRIGHT. Each range spans the two judges.

E.2 LABEL-FREE DIAGNOSTICS

The shared criterion parameters let the fitted model check a rubric without relevance labels. Table 8 reports six checks for each fit, except that the two TREC-DL fits lack Q_3. The first column states whether the difficulties rise from C1 to C5, as the rubric intends. The second names the criterion with the largest discrimination and its margin over the runner-up. The last column gives the shares of pool documents below the difficulty of C1 and at or above that of C4. Yen's Q_3 (Yen, 1984) tests local independence, the 2PL assumption that a document's answers to different criteria are independent given its ability. The statistic correlates the residuals of two criteria over documents, with $|Q_3| > 0.2$ as the warning sign. On NanoBEIR with QWEN3.5-397B, the raw correlations between criteria average 0.60 in absolute value, whereas the mean $|Q_3|$ is 0.123.

On TREC-DL, the difficulties do not rise in rubric order. C3 is easier to pass than C2, and C5 is far harder, with difficulty 7.27 under QWEN3.6-27B. Across the fits, the ratio of the 95th to the 5th

Table 8: Label-free diagnostics of each fit. Efficiency is the mean test information (Equation (8)) at each query's ideal top 10 or at the rerankers' top-10 lists, divided by the peak information. The $|Q_3|$ column gives the mean over the ten criterion pairs and, in parentheses, the number of pairs with $|Q_3| > 0.2$. TREC-DL values cover the evaluated queries of both years, and BRIGHT pools still contained the documents that BRIGHT's official evaluation excludes (Appendix C.5).

	β_c rising	Largest γ_c	τ_j p5–p95 (ratio)	Efficiency ideal top 10 / reranker top 10	$\lvert Q_3 \rvert$ (pairs)	Pool (%) $< \beta_{C1}$ / $\geq \beta_{C4}$
NanoBEIR						
QWEN3.5-397B	yes	C1 (+0.012)	0.86–2.72 (3.2)	0.66 / 0.47	0.123 (3)	79.3 / 5.1
GPT-OSS-120B	yes	C1 (+0.101)	0.95–3.18 (3.3)	0.52 / 0.33	0.151 (2)	86.7 / 3.8
BRIGHT						
QWEN3.5-397B	yes	C4 (+0.160)	0.88–2.66 (3.0)	0.70 / 0.47	0.108 (2)	73.1 / 0.9
GPT-OSS-120B	yes	C1 (+0.025)	1.04–3.17 (3.1)	0.75 / 0.45	0.111 (2)	83.2 / 1.0
ViDoRe v3						
QWEN3.5-397B	yes	C4 (+0.270)	0.77–2.29 (3.0)	0.74 / 0.60	0.112 (2)	63.9 / 2.2
TREC-DL						
QWEN3.6-27B	no	C4 (+0.108)	1.25–3.35 (2.7)	0.35 / 0.51	–	58.0 / 14.8
GPT-OSS-120B	no	C4 (+0.070)	1.45–3.65 (2.5)	0.45 / 0.55	–	67.5 / 11.1

percentile of τ_j lies between 2.5 and 3.3. Since this spread contains estimation noise, Appendix E.3 tests per-query scaling against human grades. Information efficiency divides the mean test information at a query's top 10 by the peak information of the rubric. A value of 1 means that the rubric is most precise at the top 10. At the ideal top 10, the efficiency is highest on ViDoRe v3 and BRIGHT and lowest on TREC-DL. On TREC-DL, the rubric is most informative below the documents that fill the top 10. On TREC-DL, a search over rubric subsets finds that dropping C2 raises the efficiency (Appendix G.6). On ViDoRe v3, the flagged Q_3 pairs are C1 with C2 and C4 with C5. A rubric can also pass every check and still miss aspects of relevance, which the human and NIST tests of Section 4.5 cover.

E.3 Calibration against human grades and simpler maps

This subsection supports the calibration checks of Section 4.2, using the human grades of Appendix D. Because nDCG divides by each query's ideal DCG, calibration changes RCP-nDCG only through the ratios between the nonlinear gains of a query's documents. In our human study, scoring a contest's two lists with this gain on the uncalibrated BT scores reverses the winner in 7 of the 289 contests with a verdict.

The gain against single human grades. We ask whether the gain g behaves like the probability that one annotator rates a document Useful or better. We measure this with the expected calibration error (ECE) of Section 4.2 over the study's 7,080 single grades. Over all three suites, the ECE is 0.054 (95% CI [0.042, 0.071]), where 0 means perfect calibration. At other thresholds, the ECE is higher, 0.167 against Topical or above and 0.187 against Answers or above. Of the three thresholds, the gain thus reads best as the chance of a Useful or better rating. Since no human grade enters the fit, this correspondence is an empirical result.

The judge's own answers. We check the 2PL's pass probabilities against the judge's own Stage B answers. As a simpler map, we fit a per-criterion logistic regression on the raw BT score. On NanoBEIR, the logistic regression has an ECE of 0.007, against 0.017 for the 2PL. We reran this comparison on the 2,419 native-language questions of ViDoRe v3 with judge QWEN3.5-397B. On ViDoRe v3, a sigmoid of the raw BT score is miscalibrated, with an ECE of 0.332 over 20 bins. With all queries pooled, the logistic regression and the 2PL are both well calibrated, with ECEs of 0.012 and 0.013. We also compare per-query Brier scores, the mean squared error of the predicted pass probabilities on a query's answers (lower is better). On ViDoRe v3, the Brier score is 0.0225 for the 2PL against 0.0426 for the logistic regression. The 2PL has the lower Brier score on 94.9% of queries (95% CI [94.0, 95.7]%). On NanoBEIR, the Brier score is lower for the 2PL than for the logistic regression on 97.8% of queries. Against always predicting the global pass rate, the Brier skill score of the 2PL is 0.814, where 0 means no improvement and 1 is perfect. Of the baseline Brier score, criterion base rates remove 7.6%, the BT signal 33.2%, and the per-query τ_j and α_j of the 2PL 40.6%. With GPT-OSS-120B, the Brier skill score is 0.789.

Table 9: Three scores against blind human grades, pooled over each suite's graded documents. Spearman ρ with the mean grade and AUC at Useful (0.5 is chance). z(BT) standardizes raw BT scores per query. Judge QWEN3.5-397B.

	Spearman ρ with mean grade			AUC at Useful		
Score	NanoBEIR	BRIGHT	ViDoRe v3	NanoBEIR	BRIGHT	ViDoRe v3
Raw BT $\hat{\theta}^{\mathrm{BT}}$	0.641	0.681	0.693	0.829	0.856	0.831
Per-query z(BT)	0.515	0.646	0.648	0.756	0.833	0.810
Calibrated $\tilde{\theta}$ (RCP)	0.830	0.753	0.760	0.939	0.889	0.881

Human grades across queries. Table 9 compares raw BT scores, per-query z-scores, and the calibrated $\tilde{\theta}$ of RCP, pooled over the graded documents of each study suite. Per-query standardization lowers both the Spearman correlation and the AUC below raw BT in every suite. A rescaling that ignores the rubric thus does not supply a scale shared across queries. Calibration raises the Spearman correlation with the documents' mean grade over raw BT in every suite. The increases are 0.190 on NanoBEIR (95% CI [0.122, 0.259]), 0.072 on BRIGHT (95% CI [0.026, 0.120]), and 0.067 on ViDoRe v3 (95% CI [0.005, 0.139]). Since annotators differ in leniency, we also compare same-suite pairs of documents from different queries that the same annotator graded differently. Calibration raises the share of these pairs that the score orders like the annotator's grades above that of raw BT. The increase is 0.047 (95% CI [0.035, 0.061], clustered by annotator).

E.4 Placement against NIST grades and ViDoRe qrels

This subsection supports the NIST checks of Section 4.2, repeats them on the ViDoRe v3 qrels, and tests whether the tournament's order reproduces across translations.

NIST grades on TREC-DL. Within a query, the concordance of RCP with the NIST grades is the share of differently graded passage pairs that RCP orders as NIST does (0.5 is chance). This concordance reflects Stage A alone (Proposition 3.1) and is 0.874 in 2019 and 0.828 in 2020 (Section 4.2).

The pooled Spearman correlation of $\tilde{\theta}$ with the NIST grade over all graded passages tests whether scores place passages of different queries on one scale. The ceiling reassigns each query's values of $\tilde{\theta}$ in the order of the NIST grades, and the floor shuffles them at random within each query. The observed correlation is 0.649 in 2019 and 0.595 in 2020, against ceilings of 0.768 and 0.743 and floors of 0.131 and 0.169. On graded passages that received at least one Stage B placement, calibration raises this correlation over raw BT. The increase is $+0.031$ in 2019, with an interval that includes zero, and $+0.054$ in 2020 (95% CI [0.015, 0.088]).

ViDoRe qrels. The ViDoRe v3 qrels also test the within-question order from Stage A alone. Of the same-question pairs of a grade-1 page and an unjudged pool page, RCP places the grade-1 page higher in 0.871 (0.5 is chance). For pairs of a grade-2 page and an unjudged page, the share is 0.941. Placement across questions is the part that calibration can change. With all native questions pooled, the AUC for a qrel grade of at least 1 is 0.917 for the RCP gain, against 0.892 for raw BT. The improvement over raw BT is $+0.025$ (95% CI [0.019, 0.031]). Appendix G.7 tests the tournament's order on the ViDoRe v3 pages that Count-nDCG ties.

Reproducibility across translations. We reran both stages on each of the six parallel language versions of a ViDoRe v3 question. The tournament's order reproduces across the 15 pairs of versions at a Spearman correlation of 0.962 (95% CI [0.961, 0.964]), against 0.815 for Count. A pooled Count that uses the Stage B placements of three language versions has about as many judgments per page as RCP. This pooled Count still reproduces at only 0.888, 0.074 below RCP (95% CI [0.070, 0.078]). The higher reproducibility of the tournament's order is thus not an artifact of the judging budget. A reproducible order can still be wrong, because a judge that misreads a question in every language would reproduce its error.

E.5 GAIN SHAPE

Any increasing map from $\tilde{\theta}$ to $[0, 1]$ keeps each query's order (Proposition 3.1). Such a map can therefore move a leaderboard only through the ratios between a query's gains. We compare the gain of Equation (4) with five alternatives. Three replace the discrimination weights: the unweighted mean of the pass probabilities, the logistic $\sigma(\tilde{\theta})$ centered at zero difficulty, and the C4 pass probability. Two linear maps rescale $\tilde{\theta}$, one clipped to $[0, 1]$ and one per-query min-max. The min-max map ignores calibration and equals SABER-Math's per-query rescaling under a linear gain (Georgiev et al., 2026).

On NanoBEIR and BRIGHT, the 14 rerankers' leaderboards under any alternative gain correlate with the paper's at a Spearman ρ of at least 0.859. Leaderboards thus change little, but the share of separated pairs depends more. On NanoBEIR with QWEN3.5-397B, the per-query min-max gain separates 68.1% of the pairs against 63.2% for the paper's gain. The unweighted gain separates about as many pairs. The discrimination weights thus matter little for resolution. Since the min-max gain ignores calibration yet separates more pairs, we read the share of separated pairs as resolution, not validity. We keep the discrimination-weighted gain for its probability reading (Appendix E.3) and its test-information slope (Appendix I).

F AGREEMENT WITH GROUND TRUTH AND HUMAN LABELS

This section details the label comparisons of Section 4.3: single documents against human grades, Study 1 on sets of documents, and the multi-hop queries of HotpotQA.

F.1 THE LABEL LAYER

This subsection supports the external-annotator results of Section 4.3. The label layer of the human study (Appendix D) compares each document's mean grade with its qrel and its RCP gain. The unit is a query-document pair. An annotator who graded a pair twice counts once, with the mean of the two grades. Figure 9 shows the pooled results, and Table 10 breaks them down by suite.

With the three suites pooled, the AUC at Useful (0.5 is chance) is 0.651 for the qrels, and 0.910 for the RCP gain. The gain exceeds the qrels in every suite. Among documents with qrel 0 alone, the RCP gain still separates useful from other documents with an AUC of 0.914 (95% CI [0.894, 0.933]). For same-query pairs of a useful and a non-useful document where the qrels mark only the non-useful one positive, RCP ranks the useful one higher with probability 0.791. The RCP gain here is the one stored with the study pools. Within each suite, the calibrated score of the final fit reaches the same AUC. The comparison thus does not depend on which fit supplies the scores.

Of the documents with qrel 0, 27.8% have a mean grade of at least Useful. Of the documents with a positive qrel, 37.5% fall below Useful. On NanoBEIR and BRIGHT, qrel 0 mostly means unjudged. Since a document on the topic can still correctly carry qrel 0, these shares measure disagreement with the annotators rather than qrel errors. When RCP flags the same share of documents as relevant (28.7%), it recovers 64.1% of the useful documents against 47.5% for the qrels. The false-positive rates are 7.3% for RCP against 17.3% for the qrels.

With the three suites pooled, the Spearman correlation of $\tilde{\theta}$ with the panel's mean grade is 0.792. We compare it with the panel's reliability, the expected correlation between the mean grades of two independent panels of the same size. We estimate it from random half-panels and step the correlation up with the Spearman–Brown formula, which gives 0.773. The calibrated score thus tracks the panel's mean grade about as closely as a second panel would.

Inspired by the model-based label audit of Land & Bikel (2026), we also let the ten study rerankers vote on each pooled document. The vote and the RCP gain disagree on a document when exactly one of them reaches one half. On the 760 such documents, annotators side with the gain in 82.9% (95% CI [79.1, 86.2]%). On documents whose qrel the annotators' grades contradict, the zerank rerankers agree with the qrel less often than the other rerankers (difference -7.7 pp, 95% CI $[-13.1, -1.8]$ pp).

For each of the four annotation questions, we fit a logistic curve of the annotators' yes answers on the calibrated score $\tilde{\theta}$. The threshold where this curve crosses one half rises from Topical to Complete in every suite. The threshold of Useful lies between the difficulties β_c of C2 and C3, that of Answers

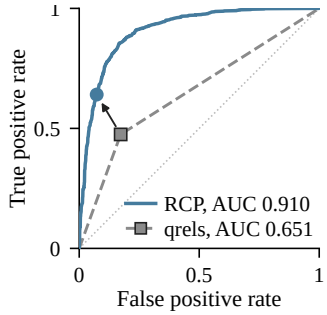


Figure 9: Documents against blind human grades (judge QWEN3.5-397B): ROC curves for a mean grade of Useful or better over 2,268 documents. The binary qrels give one operating point (square), and the arrow leads to RCP flagging the same share of documents (circle).

near C4 (on BRIGHT between C4 and C5), and that of Complete near C5. On BRIGHT, the threshold of Complete lies far above C5.

Table 10: Label layer by suite. AUCs at Useful (a mean grade of at least 2; one half is chance). The RCP gain is the one stored with the study pools. The row marked final fit comes from the final runs of Appendices C.5 and C.6. The matched operating point lets RCP flag as many documents as the qrels do. TPR is the share of useful documents flagged, and FPR the share of other documents flagged. Agreement with one annotator uses a gain of one half and a grade of 2 as thresholds. Concordant and disagreement refer to the sampling stratum of a document's contest. ViDoRe v3 qrels count any positive grade as relevant.

	NanoBEIR	BRIGHT	ViDoRe v3	All
Graded documents	723	819	726	2,268
Documents with a positive qrel (%)	28.5	19.7	39.0	28.7
AUC: qrels	0.670	0.596	0.664	0.651
RCP gain	0.939	0.889	0.881	0.910
RCP score $\tilde{\theta}$, final fit	0.939	0.889	0.881	0.906
RCP gain among qrel-0 documents	0.946	0.881	0.907	0.914
qrels / RCP gain, concordant contests	0.749 / 0.966	0.672 / 0.916	0.721 / 0.888	0.722 / 0.932
qrels / RCP gain, disagreement contests	0.635 / 0.924	0.567 / 0.878	0.641 / 0.877	0.622 / 0.900
qrel 0, mean grade at least Useful (%)	33.1	22.8	28.9	27.8
positive qrel, mean grade below Useful (%)	25.7	52.8	37.5	37.5
Matched point, TPR / FPR (%): qrels	47.2 / 13.3	33.6 / 14.3	58.0 / 25.2	47.5 / 17.3
RCP	58.0 / 4.5	56.6 / 5.6	72.8 / 14.5	64.1 / 7.3
Agreement with one annotator (%): judge	82.3	77.9	74.9	78.4
qrels	64.5	65.0	65.5	65.0
a second annotator	77.2	73.8	71.5	74.2

F.2 STUDY 1: QREL SETS AGAINST RCP SETS

This subsection reports Study 1, a comparison of qrel sets with tournament-selected sets. Study 1 asks whether the qrels or the LLM judge select the better relevant documents for a query. For a NanoBEIR query, the qrel set holds its positive documents. The RCP set holds equally many documents with the highest RCP gain, which by Proposition 3.1 are the tournament's top documents. The cases are the 264 NanoBEIR queries, of 649 in total, that have 1 to 10 positive qrels and whose two sets differ.

Three LLM judges from other model families, GLM-5.3-FLASH, DEEPSEEK-4.1-FLASH, and KIMI K3, compare the two sets blind to their source (Appendix H.4). A case counts as decided when at least two judges pick the same set in both display orders. By majority, the judges prefer the RCP set on 77.9% of the 222 decided cases (173 of 222; 95% CI [72.1, 83.3]%). On the 167 cases that all three decide the same way, the share is 82.6%. The RCP set wins most decided cases on 12 of

the 13 datasets. HotpotQA is the exception (Appendix F.3). The result holds for sets selected with GPT-OSS-120B as judge (72.2% of 216 decided cases).

On BRIGHT, the judges prefer the RCP set on 87.8% of 843 decided cases over all twelve tasks (95% CI [85.5, 89.9]%). On the 694 cases that all three decide the same way, the share is 89.6%. On AoPS, LeetCode, and TheoremQA Questions, BRIGHT's evaluation removes some listed documents, such as the query's own problem, from the rankings (Appendix C.5). The original sets for these tasks kept them. We therefore rebuilt the RCP sets with these documents removed before taking the top documents. Of the 456 original cases on these tasks, 373 remain. We dropped 72 cases whose two sets became identical. We also dropped 11 TheoremQA Questions cases whose query is an annotator's note instead of a problem. On the rebuilt cases alone, the share is 85.8% of 344. For sets selected with GPT-OSS-120B, we report only the nine tasks other than AoPS, LeetCode, and TheoremQA Questions, with 84.0% of 514 decided cases.

Two of the authors, both IR researchers, judged all 264 cases blind to which set was which, with ties allowed. As authors, they are not independent of the method, but they never saw which set came from which source. Of the 416 non-tie judgments of these two authors, 77.9% prefer the RCP set (95% Wilson CI [73.7, 81.6]%). On the 134 cases where both of them prefer the same set, the RCP set wins 120 (89.6%, 95% CI [83.2, 93.7]%).

Study 1 conditions on queries whose two sets differ and does not compare the label sources on all documents. The external annotators' grades cover all 2,268 documents of the human study, whether or not the qrels and RCP agree on them (Section 4.3 and Appendix F.1).

F.3 Multi-hop queries: HotpotQA

NanoHotpotQA is the dataset on which Studies 1 and 2 (Appendices F.2 and H.4) went most strongly against RCP. Its questions need a primary and a secondary fact, each from its own clearly relevant document, and its qrels mark exactly these two documents. RCP and our annotators judge each document on its own and do not penalize redundancy. A second document on the primary fact can therefore outrank the document with the secondary fact, which the qrel set contains by construction. In Study 1, the judges preferred the qrel set in all 8 decided HotpotQA cases. The authors who judged Study 1 gave the RCP set 0 of their 12 non-tie judgments. In Study 2, the list of RCP-nDCG won 4 of 14 decided HotpotQA sign-flips.

In the external human study, only 6 HotpotQA contests enter the main measure, too few for a claim. In 3 of them, the majority verdict sides with RCP-nDCG. Relevance that arises only from combining documents would need criteria or comparisons that look at several documents at once. RCP has neither, nor does it model diversity or the needs of individual users.

G Ranking Results

This section details the ranking results of Section 4.4: failure modes of nDCG@10, sensitivity and scores per dataset, second judges, the zerank rerankers, robustness, and Count-nDCG.

G.1 Failure modes of nDCG@10

On a query, nDCG@10 *saturates* when all 14 rerankers score 1 and *floors* when all 14 score 0. The metric *compresses* the rerankers when neither holds and at least half of the 91 reranker pairs tie, that is, differ by at most 10^{-6}. Figure 10 gives the share of queries with at least one mode on NanoBEIR, BRIGHT, and TREC-DL. At least one mode occurs on 45.0% of the NanoBEIR queries and on 30.7% of the BRIGHT queries. On ViDoRe v3, the share is 23.1%. With the graded NIST labels of TREC-DL, it is 7.0% in 2019 and 5.6% in 2020.

G.2 Sensitivity per dataset

Per dataset, RCP-nDCG separates more pairs than qrel-nDCG on 11 of the 13 NanoBEIR datasets and on all 12 BRIGHT tasks. The exceptions are ArguAna and HotpotQA, where RCP-nDCG separates fewer pairs than qrel-nDCG. Appendix F.3 discusses HotpotQA.

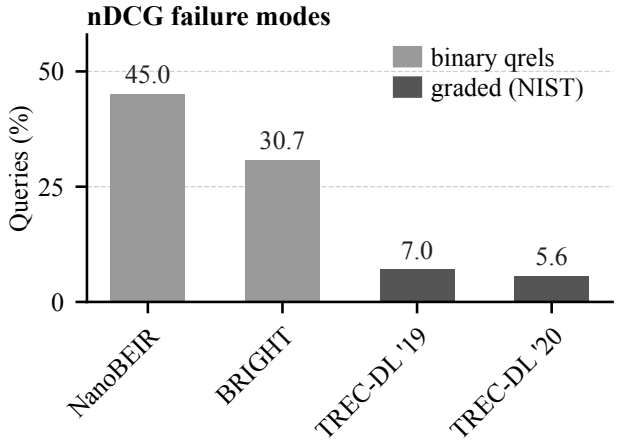


Figure 10: Share of queries on which nDCG@10 saturates, floors, or compresses the rerankers. With the graded NIST labels of TREC-DL, nDCG@10 rarely shows any of these modes.

G.3 Reranker scores per dataset

Table 11 lists nDCG@10 on the qrels and RCP-nDCG@10 for every reranker and NanoBEIR dataset. Tables 12 and 13 do the same for the BRIGHT tasks and the ViDoRe v3 corpora. Table 14 adds the suite means on NanoBEIR and BRIGHT under both judges. Since the two metrics differ in level, only their orders of the rerankers are comparable.

G.4 Second judges

A metric built on one LLM judge should change little with another. We therefore ran both stages with GPT-OSS-120B (OpenAI, 2025) on NanoBEIR and BRIGHT, and on the TREC-DL pools, where the calibration CAL-C fits its judgments (Appendix C.7). Table 14 gives the suite means of qrel-nDCG@10 and Count-nDCG@10, and of RCP-nDCG@10 under both judges. Within each dataset, we correlate the two judges' RCP-nDCG@10 over the 14 rerankers, where 1 means identical leaderboards. The mean Spearman correlation over datasets is 0.958 on NanoBEIR and 0.988 on BRIGHT. No dataset falls below 0.916. On TREC-DL, the two judges' leaderboards of all 18 systems correlate at Kendall 0.987 with both years pooled. The two judges' calibrated scores over the 27,875 pool passages correlate at Spearman 0.934. Under the second judge, RCP-nDCG@10 separates 60.5% of the reranker pairs on NanoBEIR, against 63.2% under the primary judge (Appendix G.2).

On TREC-DL, the within-query concordance is the share of differently graded passage pairs that RCP orders as NIST does, with 0.5 at chance. The concordance is 0.859 (2019) and 0.822 (2020) with GPT-OSS-120B, against 0.874 and 0.828 with QWEN3.6-27B. A third model, GLM-5.2, arbitrated the 582 TREC-DL passage pairs on which the two judges disagree most. On the 148 pairs that NIST grades differently, GLM-5.2 orders 69.6% as NIST does, against 60.1% for QWEN3.6-27B. Pooling the two judges' Stage B answers in one fit does not improve on the better single judge (Appendix E.4).

G.5 The zerank rerankers

The three zerank rerankers learn from pairwise preferences of an ensemble of LLMs (Pipitone et al., 2025). A metric built on an LLM judge could therefore favor them (Section 4.4). On NanoBEIR, qrel-nDCG@10 ranks them last, while RCP-nDCG@10 ranks zerank-2 and zerank-1 first and second. The mean RCP-nDCG@10 of the three zerank rerankers exceeds that of the other 11 rerankers by 5.4 pp over all datasets and on 11 of the 13 datasets. On BRIGHT, qrel-nDCG@10 already ranks the two first and second (Table 14).

On queries with at least five qrel-positive documents, we correlate each reranker's scores of these documents with their calibrated scores (Spearman, averaged over queries). The three zerank rerankers reach 0.79 on NanoBEIR and 0.63 on BRIGHT, against 0.63 and 0.46 for the other 11 rerankers. A

Table 11: nDCG@10 with the benchmark's qrels and RCP-nDCG@10 on NanoBEIR (judge QWEN3.5-397B), in %. Each block is sorted by its mean over the 13 datasets. Arrows give each reranker's change in rank from nDCG@10 to RCP-nDCG@10. Bold marks the best reranker per column. Colors mark the model family: Cohere, Contextual AI, Qwen3, Voyage AI, ZeroEntropy, and Jina AI.

	Datasets													Mean
Reranker	ArguAna	ClimateFever	DBPedia	FEVER	FiQA	HotpotQA	MSMARCO	NFCorpus	NQ	Quora	SCIDOCS	SciFact	Touché2020	
nDCG@10														
Contextual v2 6B	**83.3**	42.1	77.1	**98.7**	68.3	**98.0**	69.6	48.6	**84.4**	89.4	50.8	**90.3**	62.5	**74.1**
Qwen3-Reranker-8B	77.8	**54.1**	75.0	96.2	69.0	94.3	72.3	46.7	81.6	97.1	48.9	83.6	65.9	74.0
Qwen3-Reranker-4B	71.5	51.7	74.6	97.3	67.1	94.3	70.4	47.4	82.1	96.0	51.5	82.9	**66.6**	73.3
Rerank 4 Pro	66.7	52.3	75.8	97.8	**70.8**	94.4	67.2	47.9	81.7	96.4	46.4	81.7	63.6	72.5
Contextual v2 2B	67.0	41.1	75.9	95.2	65.1	96.3	68.3	48.2	80.1	93.2	**51.7**	84.4	62.4	71.5
Voyage 2.5	80.9	37.6	**77.6**	96.8	67.6	95.0	70.2	46.6	80.9	93.3	44.2	82.3	55.6	71.4
Rerank 4 Fast	63.3	52.3	74.3	96.9	61.5	93.7	**72.6**	46.9	79.2	95.4	44.6	80.4	61.8	71.0
Voyage 2.5 Lite	76.9	39.4	76.5	94.8	69.5	93.9	70.7	45.2	77.6	94.1	42.4	82.0	57.8	70.8
Contextual v2 1B	60.9	40.0	73.7	96.0	62.4	95.2	68.2	46.9	73.4	93.7	47.1	83.9	61.1	69.4
Jina Reranker v3	79.3	39.6	68.0	97.7	57.9	96.3	69.5	**49.0**	79.2	**97.9**	34.3	78.5	52.9	69.2
Qwen3-Reranker-0.6B	67.0	47.8	70.6	96.5	58.6	92.0	67.9	46.1	75.3	93.6	44.9	80.4	58.7	69.2
zerank-1-small	55.3	33.5	73.8	91.7	62.2	89.4	62.8	46.4	74.8	82.8	44.1	82.2	65.3	66.5
zerank-1	53.5	33.7	73.6	92.3	63.6	90.2	62.0	48.1	77.9	78.6	42.9	82.9	64.3	66.4
zerank-2	49.6	36.5	74.0	92.1	62.9	90.1	59.2	47.1	77.0	76.8	45.3	83.2	64.0	66.0
RCP-nDCG@10														
zerank-2 ▲13	51.8	**88.9**	**90.6**	**93.4**	90.8	89.5	**88.3**	**87.9**	83.9	86.5	**90.4**	**91.1**	**97.0**	**86.9**
zerank-1 ▲11	52.4	87.3	**90.6**	93.0	**91.3**	88.9	88.0	87.5	84.2	88.1	89.8	90.3	96.8	86.8
Voyage 2.5 ▲3	63.1	85.4	89.9	92.8	83.1	**92.0**	85.8	84.8	81.8	87.4	79.6	87.0	87.1	84.6
zerank-1-small ▲8	50.7	83.7	87.1	91.3	90.0	88.0	85.8	81.8	79.3	83.8	**90.4**	88.5	96.1	84.3
Rerank 4 Pro ▼1	57.6	71.1	86.9	85.7	86.1	90.3	85.9	85.2	83.0	88.1	80.4	88.0	90.2	83.0
Contextual v2 6B ▼5	**66.1**	68.2	88.4	84.8	80.3	90.1	87.9	83.5	**85.4**	78.9	84.7	84.9	91.4	82.7
Voyage 2.5 Lite ▲1	61.0	82.4	87.7	91.7	79.7	90.3	84.3	79.0	77.7	81.1	84.7	84.4	83.0	82.1
Qwen3-Reranker-4B ▼5	59.5	68.7	84.5	88.2	84.2	89.7	83.3	81.6	80.9	87.6	86.4	85.4	86.1	82.0
Qwen3-Reranker-8B ▼7	63.6	64.7	86.6	87.7	85.3	89.7	83.6	82.4	81.3	**88.3**	79.1	85.7	87.6	82.0
Contextual v2 2B ▼5	52.5	70.7	87.5	85.4	81.1	89.2	87.2	81.9	83.7	81.2	88.8	85.2	90.6	81.9
Rerank 4 Fast ▼4	56.5	64.7	83.7	85.4	83.0	89.1	83.1	79.3	77.8	85.9	84.6	86.1	89.9	80.7
Contextual v2 1B ▼3	52.5	67.0	86.1	85.4	81.9	87.0	86.0	79.6	78.1	84.8	84.6	83.5	89.3	80.5
Qwen3-Reranker-0.6B ▼2	54.7	73.7	78.3	87.3	77.0	87.0	80.2	73.4	71.6	83.9	84.2	83.3	78.7	77.9
Jina Reranker v3 ▼4	52.8	64.9	73.3	79.4	70.0	84.4	81.6	62.0	69.1	75.1	44.2	69.8	77.4	69.5

preference shared with the judge would also produce this lead. This result therefore locates the lead but cannot tell whether it reflects better rankings.

On TREC-DL, RCP-nDCG ranks zerank-2 first among the 18 systems in both years, whereas NIST ranks it 14th and 9th. NIST counts an unjudged passage as irrelevant. Unjudged passages fill 8.9% of zerank-2's top-10 slots, against 3.7–5.4% for the 9 systems from other families that NIST ranks above it with both years pooled. Condensed lists (Sakai, 2007) drop every unjudged passage before scoring. On condensed lists, the Kendall correlation between the leaderboards of RCP-nDCG and NIST closes 0.538 of its gap to 1 under the linear gain (CAL-A) and 0.348 under exponential gain (CAL-B). On condensed lists, zerank-2 ranks 9th in 2019 and 5th in 2020, against 14th and 9th on the full lists. On the full lists, it trails the NIST leader by 0.048 in 2019 (95% CI [-0.006, 0.102]) and 0.030 in 2020 ([-0.012, 0.071]).

In the human study on NanoBEIR, BRIGHT, and ViDoRe v3, annotators never saw a system name. In the 153 contests that pit a zerank list against another reranker's list and reach a verdict, the verdict picks the zerank list in 69.3% (95% CI [61.9, 76.6]%). Where RCP-nDCG picks the zerank list and qrel-nDCG the other list, the verdict picks the zerank list in 78.8% of 66 contests ([68.8, 88.8]%).

Table 12: nDCG@10 with the benchmark's qrels and RCP-nDCG@10 on BRIGHT (judge QWEN3.5-397B), in %, with the documents that BRIGHT excludes for a query removed. Each block is sorted by its mean over the 12 tasks. Arrows give each reranker's change in rank from nDCG@10 to RCP-nDCG@10. Bold marks the best reranker per column. Colors mark the model family as in Table 11.

	Datasets												Mean
Reranker	AoPS	Biology	Earth Sci.	Economics	LeetCode	Pony	Psychology	Robotics	StackOvfl.	Sust. Liv.	ThQA-Q	ThQA-T	
						nDCG@10							
zerank-1	14.4	**54.6**	51.7	33.6	21.0	38.7	**47.6**	33.2	27.6	44.0	46.2	**46.9**	**38.3**
zerank-2	13.5	52.9	**52.0**	**37.4**	22.1	25.2	46.5	**37.7**	**31.0**	**44.9**	46.3	43.6	37.8
Voyage 2.5	10.2	47.9	49.6	30.9	21.1	**41.6**	42.3	30.9	26.0	38.9	37.3	30.0	33.9
Rerank 4 Pro	12.7	50.2	49.2	31.0	29.2	3.8	44.5	27.8	24.2	35.8	**48.1**	45.6	33.5
zerank-1-small	10.4	46.5	45.6	29.1	19.2	24.4	41.0	25.5	26.6	40.1	45.7	40.2	32.9
Rerank 4 Fast	10.9	44.6	45.5	30.8	20.4	3.9	38.2	23.6	23.7	34.2	46.7	38.8	30.1
Voyage 2.5 Lite	11.1	38.4	45.3	26.6	24.9	26.2	38.0	23.5	22.2	30.8	33.3	21.9	28.5
Qwen3-Reranker-4B	12.2	30.7	36.6	21.7	**40.7**	18.7	27.9	23.1	26.2	23.9	42.8	36.3	28.4
Qwen3-Reranker-8B	9.4	32.7	34.4	25.6	29.4	21.2	30.7	27.4	24.4	25.4	39.7	34.1	27.9
Jina Reranker v3	8.4	26.8	39.2	27.8	10.2	18.1	29.2	23.2	12.5	31.5	28.3	32.6	24.0
Qwen3-Reranker-0.6B	8.7	19.1	24.6	14.9	28.9	3.1	17.1	13.8	15.4	11.3	37.1	30.0	18.7
Contextual v2 6B	**17.4**	10.0	16.7	18.0	15.0	29.2	20.7	12.3	5.7	11.5	38.2	22.8	18.1
Contextual v2 1B	11.4	10.2	24.1	18.6	21.2	16.3	19.5	13.9	10.8	12.3	37.2	18.7	17.9
Contextual v2 2B	14.5	11.6	23.7	16.0	14.7	12.5	23.3	9.5	6.5	12.4	34.0	28.7	17.3
						RCP-nDCG@10							
zerank-2 ▲1	**75.5**	**85.9**	**86.5**	**87.8**	77.0	**73.3**	**82.2**	**85.2**	**82.3**	**84.0**	**81.3**	77.8	**81.6**
zerank-1 ▼1	70.1	84.5	85.4	86.1	75.9	68.7	81.8	82.4	80.4	83.6	80.7	76.4	79.7
zerank-1-small ▲2	71.4	76.8	81.9	80.5	70.1	52.7	76.5	77.3	74.4	77.8	76.3	70.8	73.9
Voyage 2.5 ▼1	54.4	78.7	83.2	78.8	69.1	63.4	79.1	78.7	72.2	79.1	69.6	59.4	72.1
Rerank 4 Pro ▼1	73.1	76.9	81.6	76.8	**78.7**	15.5	76.6	77.5	67.7	72.7	75.2	**79.1**	71.0
Qwen3-Reranker-4B ▲2	68.8	68.4	74.3	65.4	70.1	63.1	61.9	70.4	72.7	60.2	72.2	70.4	68.1
Qwen3-Reranker-8B ▲2	64.7	69.7	73.3	70.3	69.4	33.8	65.8	74.4	71.7	65.8	70.8	68.6	66.5
Rerank 4 Fast ▼2	66.8	73.9	78.5	74.3	67.2	19.3	71.3	71.7	65.7	71.9	70.1	67.2	66.5
Voyage 2.5 Lite ▼2	45.3	71.4	78.5	73.3	62.4	47.1	70.8	72.5	64.3	71.4	63.0	49.1	64.1
Qwen3-Reranker-0.6B ▲1	59.6	53.3	61.1	55.1	61.3	33.7	48.4	61.9	52.9	48.9	63.2	57.3	54.7
Jina Reranker v3 ▼1	41.1	59.1	65.1	62.4	50.9	40.5	55.4	68.1	46.9	62.2	46.4	54.3	54.4
Contextual v2 1B ▲1	47.5	41.7	48.9	48.4	32.0	30.0	50.0	49.4	31.1	38.6	51.7	44.7	42.8
Contextual v2 2B ▲1	36.5	38.2	46.5	45.6	42.3	21.1	50.0	49.0	31.8	34.9	53.8	51.1	41.7
Contextual v2 6B ▼2	38.5	40.1	45.4	45.5	36.9	30.5	50.5	47.1	27.9	32.5	52.7	50.5	41.5

Where RCP-nDCG picks the zerank list with a margin of at least 0.02, the verdict picks it in 85.3% of 102 contests ([78.4, 92.2]%).

On ViDoRe v3, zerank-2 and zerank-1 rise from qrel ranks 4 and 5 to 1 and 2. The rise already appears under Count-nDCG, which scores documents by the rubric alone.

G.6 RUBRIC, PROMPT, AND SIMULATION ROBUSTNESS

This subsection tests how the leaderboard of RCP-nDCG depends on the rubric, the Stage A prompt, and the criteria, and checks the gain in a simulation.

Rubric and prompt factorial. On TREC-DL with QWEN3.6-27B, we cross five Stage A prompts with sixteen rubrics and refit the 2PL in each of the 80 cells (Table 15). The four other Stage A prompts keep the relevance framing and score anchors of the paper's prompt and replace only its list of scoring criteria. Two paraphrase this list, and two swap in answer-quality or document-quality criteria. Their tournament orders stay close to the shipped one, with mean per-query Spearman correlations of 0.968 to 0.971. The four prompts therefore test the wording of the criteria list rather than a different construct. A pair counts as separated when a paired percentile bootstrap's 95%

Table 13: nDCG@10 with the benchmark's graded qrels as linear gains and RCP-nDCG@10 on ViDoRe v3 (judge QWEN3.5-397B), in %, on the native-language questions. Tied reranker scores receive their group's mean gain. Each block is sorted by its mean over the 8 corpora. Arrows give each reranker's change in rank from nDCG@10 to RCP-nDCG@10. Bold marks the best reranker per column. Colors mark the model family as in Table 11.

	Datasets								Mean
Reranker	Comp. Sci.	Energy	Finance (en)	Finance (fr)	HR	Industrial	Pharma.	Physics	
	nDCG@10								
Voyage 2.5	**85.1**	**74.0**	75.1	**58.5**	**71.9**	**62.1**	**71.6**	50.3	**68.6**
Rerank 4 Pro	83.2	73.7	**75.5**	57.9	70.8	61.3	71.4	51.5	68.2
Voyage 2.5 Lite	83.8	73.2	74.6	57.3	69.6	**62.1**	70.2	49.0	67.5
zerank-2	84.2	72.4	75.1	55.8	68.9	60.7	69.8	49.0	67.0
zerank-1	84.0	71.7	75.1	55.4	68.7	60.4	69.7	48.9	66.7
Rerank 4 Fast	81.9	70.7	74.5	53.6	68.5	59.7	69.5	48.8	65.9
Contextual v2 6B	74.4	71.8	73.8	55.2	69.1	60.6	68.2	51.2	65.5
Qwen3-Reranker-8B	80.8	70.6	70.3	55.5	65.5	60.0	68.6	**52.3**	65.4
zerank-1-small	84.3	69.5	72.8	52.6	66.8	58.7	67.5	48.5	65.1
Contextual v2 2B	76.7	71.6	71.7	53.2	68.1	61.1	67.5	49.0	64.9
Contextual v2 1B	77.7	68.9	69.1	49.7	65.4	57.4	66.4	49.2	63.0
Qwen3-Reranker-4B	80.4	67.9	67.4	51.0	61.7	57.8	66.4	51.0	62.9
Jina Reranker v3	74.1	65.8	63.6	41.5	57.2	57.3	67.9	45.3	59.1
Qwen3-Reranker-0.6B	74.4	53.3	57.3	34.2	45.2	49.7	60.0	47.9	52.8
	RCP-nDCG@10								
zerank-2 ▲3	**94.2**	**93.6**	87.6	89.6	**92.0**	**89.4**	**91.2**	**93.7**	**91.4**
zerank-1 ▲3	94.1	93.3	**88.9**	89.6	91.6	89.1	90.8	93.3	91.3
Voyage 2.5 ▼2	93.3	93.3	86.5	87.5	91.8	89.1	90.3	91.0	90.4
Rerank 4 Pro ▼2	90.8	93.0	88.6	**90.2**	90.4	87.9	89.0	91.6	90.2
zerank-1-small ▲4	93.2	90.9	88.2	86.7	90.1	87.8	88.1	90.8	89.5
Voyage 2.5 Lite ▼3	91.5	90.7	86.9	85.2	89.8	87.7	88.6	86.9	88.4
Rerank 4 Fast ▼1	89.8	90.2	87.2	84.7	88.4	85.1	86.7	88.2	87.5
Qwen3-Reranker-8B	88.3	89.2	82.6	85.1	85.3	84.1	85.9	89.2	86.2
Contextual v2 2B ▲1	82.8	89.9	85.7	83.7	87.9	85.0	85.6	88.0	86.1
Contextual v2 6B ▼3	79.0	89.0	86.6	85.5	88.0	85.1	84.6	88.0	85.7
Contextual v2 1B	85.1	88.1	82.5	82.2	85.3	81.5	83.7	86.7	84.4
Qwen3-Reranker-4B	87.7	86.5	79.5	80.4	82.7	81.0	84.1	88.1	83.7
Qwen3-Reranker-0.6B ▲1	82.5	79.8	75.1	67.2	72.5	76.4	78.3	83.2	76.9
Jina Reranker v3 ▼1	73.2	70.5	67.9	64.3	69.7	74.3	75.9	76.9	71.6

interval excludes zero. The reference cell, with the shipped prompt and rubric, separates 76 of the 91 pairs. The 59 cells whose rubric still measures relevance reverse none of these pairs, and their leaderboards correlate with the reference at Kendall 0.890 or higher. All 20 reversals among the 6,004 opportunities come from rubrics built to fail. Of these, 17 come from the all-easy rubric, 2 from the all-hard rubric, and 1 from a document-quality rubric that ignores the query. With the shipped rubric, the four other prompts also reverse no separated pair.

Criterion subsets. We refit the TREC-DL model with each non-empty subset of the five criteria. By information efficiency at the ideal top 10 (Appendix E.2), the full rubric ranks third. Its information efficiency is 0.407. The best subset drops C2 and reaches 0.446, and single criteria reach between 0.073 (C5 alone) and 0.297 (C4 alone). The subset search uses no relevance labels.

Simulation. A simulation with known truth, 200 replications of 50 queries, tests the gain and the paired t-test, not the tournament or the 2PL fit. Over the 16 combinations of pool regime and judge noise, RCP-nDCG separates 3.8–6.4% of the truly equal pairs and nDCG 3.4–6.4%. Both rates sit near the nominal 5% level of the test. With an accurate judge, RCP-nDCG detects more true differences than nDCG in every pool regime. Under graded labels, RCP-nDCG detects fewer true

Table 14: Cross-suite summary: suite means over datasets (%), with the rank among the 14 rerankers in parentheses. Columns give nDCG@10 on the qrels (qrel), Count-nDCG@10 (Count) and RCP-nDCG@10 (RCP) with judge QWEN3.5-397B, and RCP-nDCG@10 with the second judge GPT-OSS-120B (RCP_2). BRIGHT removes each query's excluded documents (Appendix C.5). Rows follow RCP-nDCG@10 on NanoBEIR. Bold marks the best reranker per column.

	NanoBEIR				BRIGHT			
Reranker	qrel	Count	RCP	RCP_2	qrel	Count	RCP	RCP_2
zerank-2	66.0 (14)	**88.6** (1)	**86.9** (1)	86.7 (2)	37.8 (2)	**79.5** (1)	**81.6** (1)	**80.8** (1)
zerank-1	66.4 (13)	88.4 (2)	86.8 (2)	**86.7** (1)	**38.3** (1)	77.5 (2)	79.7 (2)	79.3 (2)
Voyage 2.5	71.4 (6)	85.6 (4)	84.6 (3)	85.1 (3)	33.9 (3)	70.1 (4)	72.1 (4)	71.0 (4)
zerank-1-small	66.5 (12)	86.0 (3)	84.3 (4)	84.6 (4)	32.9 (5)	72.4 (3)	73.9 (3)	73.4 (3)
Rerank 4 Pro	72.5 (4)	84.0 (5)	83.0 (5)	83.2 (5)	33.5 (4)	68.0 (5)	71.0 (5)	70.5 (5)
Contextual v2 6B	**74.1** (1)	83.0 (8)	82.7 (6)	82.3 (7)	18.1 (12)	39.8 (14)	41.5 (14)	37.3 (14)
Voyage 2.5 Lite	70.8 (8)	83.7 (6)	82.1 (7)	82.4 (6)	28.5 (7)	62.1 (9)	64.1 (9)	62.0 (9)
Qwen3-Reranker-4B	73.3 (3)	83.1 (7)	82.0 (8)	81.7 (10)	28.4 (8)	66.2 (6)	68.1 (6)	66.2 (6)
Qwen3-Reranker-8B	74.0 (2)	82.8 (10)	82.0 (9)	82.1 (9)	27.9 (9)	64.4 (8)	66.5 (7)	65.1 (7)
Contextual v2 2B	71.5 (5)	82.8 (9)	81.9 (10)	82.2 (8)	17.3 (14)	40.1 (13)	41.7 (13)	37.5 (13)
Rerank 4 Fast	71.0 (7)	82.1 (11)	80.7 (11)	81.1 (11)	30.1 (6)	64.8 (7)	66.5 (8)	65.0 (8)
Contextual v2 1B	69.4 (9)	81.8 (12)	80.5 (12)	80.3 (12)	17.9 (13)	41.6 (12)	42.8 (12)	37.8 (12)
Qwen3-Reranker-0.6B	69.2 (11)	79.7 (13)	77.9 (13)	77.2 (13)	18.7 (11)	53.4 (10)	54.7 (10)	52.2 (10)
Jina Reranker v3	69.2 (10)	70.4 (14)	69.5 (14)	71.2 (14)	24.0 (10)	52.5 (11)	54.4 (11)	52.0 (11)

Table 15: Rubric and prompt factorial on TREC-DL (QWEN3.6-27B). Each family crosses its rubrics with the five Stage A prompts, except the reference family, which pairs the shipped rubric with the four other prompts. Reversals count the pairs that the reference cell separates and whose sign flips, summed over the family's cells. Kendall is the rank correlation of each cell's 14-reranker leaderboard with the reference.

Family	Rubrics	Cells	Reversals	Kendall range
Criterion wording	three paraphrases of the rubric	15	0	0.934–1.000
Criterion count	one, three, or seven criteria	15	0	0.890–0.978
Criterion ordering	scrambled order	5	0	0.934–1.000
Criterion redundancy	near-duplicate or diverse facets	10	0	0.934–1.000
Stage A criteria wording	shipped rubric, four other prompts	4	0	0.956–0.978
Difficulty placement	all easy, all hard, tiled, top-heavy	20	19	0.714–1.000
Construct validity	document quality, trivial items	10	1	0.868–0.956

differences than nDCG from a judge noise of 0.50, although the judge then still ranks documents more accurately than every simulated reranker. Under binary labels, this happens only from a judge noise of 1.00, where the judge ranks documents less accurately than every simulated reranker.

G.7 RCP-nDCG AND COUNT-NDCG

Count-nDCG scores documents by the rubric alone (Equation (5)). Count-nDCG separates about as many reranker pairs as RCP-nDCG (Section 4.4). Within a query, however, the tournament orders documents that Count-nDCG ties. The tests below concern this order. Except for the per-dataset leaderboards, the BRIGHT values of this subsection come from reranker lists that still contain each query's excluded documents (Appendix C.5).

External judges. Three LLM judges from other model families compare two top-10 lists per case (Appendix H.5). The judges are blind to the metrics and see each case twice with swapped labels (Table 16). A case counts when at least two of the three external judges name the same list in both label orders. An ideal list holds a query's ten best pool documents under each gain. A minimal pair sets a reranker's top 10 against the same list with its Count-tied documents in the tournament's order. Table 17 reports the per-judge, per-suite ideal-list shares, which range from 80.6% to 98.8%. The judges favor the tournament's order in 86.8% of decided ideal lists and 84.0% of minimal pairs. The judges side with the winner of RCP-nDCG in 63.1% of the per-query cases where the two metrics name opposite winners. On controls, where both metrics name the same winner, the judges pick

Table 16: Share (%) of decided cases favoring the order of RCP-nDCG over Count-nDCG [95% CI] (n). 50 means no preference. For controls, the share counts the common winner. Pooled over NanoBEIR, BRIGHT, TREC-DL, and ViDoRe v3.

Test (unit)	LLM judges
Ideal lists (queries)	86.8 [83.1, 90.3] (333)
Minimal pairs (reranker lists)	84.0 [73.5, 93.5] (50)
Per-query disagreements (reranker pairs)	63.1 [57.3, 68.5] (317)
both differences beyond their tie zones	66.1 [57.1, 75.2] (112)
only the difference in RCP-nDCG beyond	72.7 [60.0, 84.3] (55)
only the difference in Count-nDCG beyond	52.1 [40.0, 63.8] (73)
Controls: common winner (reranker pairs)	75.0 [65.3, 84.3] (76)

Table 17: Share (%) of decided queries on which each LLM judge prefers the ideal top-10 list of RCP-nDCG over that of Count-nDCG, with 95% CIs and the number of decided queries (n); 50% means no preference. These are the values of Figure 6b. On BRIGHT, the lists can contain documents that BRIGHT's evaluation excludes (Appendix C.5).

Judge	NanoBEIR	BRIGHT	ViDoRe v3	TREC-DL
GLM-5.3-FLASH	86.2 [77.2, 94.5] $n=58$	80.8 [71.2, 89.3] $n=73$	81.4 [73.2, 89.1] $n=97$	93.6 [87.8, 98.7] $n=78$
DEEPSEEK-4.1-FLASH	86.5 [78.4, 93.4] $n=74$	84.1 [75.9, 92.0] $n=82$	80.9 [73.1, 87.9] $n=110$	94.2 [88.6, 98.8] $n=86$
KIMI K3	85.7 [76.8, 93.2] $n=70$	81.2 [72.5, 89.3] $n=80$	80.6 [72.8, 87.7] $n=108$	98.8 [96.2, 100.0] $n=82$

that winner in 75.0%. The judges can thus see differences of this size. Each metric has a tie zone, a margin up to which we read its difference as a tie (Appendix H.5). Where both metrics' differences exceed their tie zones, the judges side with RCP-nDCG in 66.1%.

Tie resolution. We take same-query document pairs that Count-nDCG ties exactly and that a reference orders, keeping pairs that can change a top-10 list. The tournament orders 0.671 of these pairs like NIST, 0.680 like the human panel, and 0.660 like the ViDoRe v3 qrels. All three accuracies are well above one half (Table 18).

DL19 references. Parry et al. (2025) re-assessed the 43 queries of DL19, and we use their grades for analysis only (Appendix K). The two re-assessors graded 90.9% of the slots of RCP's ideal top 10 and 90.2% of those of Count's. An ungraded passage counts as 0. Under nDCG@10 on its own grades, NIST alone prefers neither ideal list. The consensus, the mean grade of NIST and both re-assessors, prefers the tournament's list on 72.1% of the queries where the two lists differ (95% CI [58.1, 83.7]%, Table 19). On the 25,194 DL19 pairs that NIST separates, Count ties, and both re-assessors graded, the tournament orders 0.671 like NIST. On the pairs that a re-assessor separates, it agrees with that re-assessor on 0.757, whereas NIST agrees with the re-assessors on 0.700. Single NIST grades thus appear noisy for such fine distinctions.

Minimal pairs. For each reranker and query, we build two versions of its top 10 that differ only in the order of exactly Count-tied documents. One version follows the tournament's order and one reverses it. Unlike the minimal pairs of the LLM judges (Appendix H.5), these pairs use the reversed order, not the reranker's own order, as the alternative. Count-nDCG scores both alike. The qrels of each suite, NIST, and both re-assessors all prefer the tournament's version. Each reference's paired test over queries favors that order for all 14 rerankers of each suite, except on DL19 under NIST alone (5 of 14) and under one re-assessor (8 of 14).

Saturation and per-dataset leaderboards. Moving each reranker's top 10 toward Count's ideal list by greedy swaps raises Count's share of exact ties among the five best rerankers. On TREC-DL it rises from 15.9% to 66.8%. Elsewhere it reaches 7.8–16.5%, whereas RCP-nDCG ties at most 1.0%. Per dataset, the top-ranked reranker differs between the metrics in 10 of the 35 datasets.

Table 18: Tie resolution: accuracy on same-query document pairs that Count ties exactly and a reference orders, restricted to pairs that can change a top-10 list. Ties count one half, chance 0.5. The NIST row uses the TREC-DL run of the named judge, the other rows QWEN3.5-397B. Brackets are 95% query-cluster bootstrap intervals.

Reference	Pairs	Tournament
NIST (QWEN3.6-27B)	26,730	0.671 [0.620, 0.725]
Human panel	256	0.680 [0.608, 0.752]
ViDoRe v3 qrels	16,187	0.660 [0.630, 0.691]

Table 19: DL19 references built from NIST and the re-assessments of Parry et al. (2025) (43 queries). Ideal lists: share of queries, among those where the reference scores the two differently, where it scores the top 10 of RCP above that of Count. Ties break by passage identifier. ΔnDCG@10 is the mean difference in nDCG@10. Minimal pairs: share of reranker top-10 lists where the reference prefers the tournament's order of Count-tied passages. Rerankers gives the number (of 14) whose paired t-test over queries favors that order ($p < .05$). Brackets are 95% CIs from a query bootstrap.

	Ideal lists		Minimal pairs	
Reference	RCP wins (%)	ΔnDCG@10	tournament (%)	rerankers
NIST	48.8 [33.3, 63.4]	+0.025	62.4	5
Re-assessor 1	71.4 [57.1, 83.7]	+0.076	75.1	14
Re-assessor 2	61.9 [46.5, 76.2]	+0.059	64.5	8
Consensus (mean of all three)	72.1 [58.1, 83.7]	+0.051	72.2	14

H EXTERNAL PREFERENCE VALIDATION

H.1 NIST-DECISIVE RERANKER PAIRS

We compare each metric with the NIST grades of TREC-DL in the pool and top-150 panels of Appendix C.7. A pair is NIST-decisive when qrel-nDCG@10 with the NIST grades as linear gains separates it (Section 4.5). Table 20 counts, for each year and panel, the reranker pairs that each metric separates and the NIST-decisive pairs on which it agrees with NIST. The share of separated pairs is a metric's sensitivity (Sakai, 2006; Clarke et al., 2021) (Appendix B.2). A metric without signal would separate about 5% of pairs. In the pool panel, qrel-nDCG@10 separates 14 of the 91 pairs in 2019 and 37 in 2020. Count-nDCG and RCP-nDCG separate many more pairs in both years. Since Count-nDCG scores documents by the rubric alone, the rubric supplies most of the added resolution.

In the pool panel with linear gain, RCP-nDCG agrees with NIST on every NIST-decisive pair in both years. The agreement holds under each of its three calibration fits, CAL-A, CAL-B, and CAL-C (Appendix C.7), of which Table 20 shows the first two. Under exponential gain, in the top-150 panel, and in every other setting we test, the few NIST-decisive pairs that go the other way all involve zerank-2 or zerank-1 (Appendix G.5). When we test each pair once over the queries of both years, in the pool panel with linear gain, 39 pairs are NIST-decisive. RCP-nDCG agrees with NIST on 36 of them. RCP-nDCG separates only 1 pair in the direction opposite to NIST, Contextual v2 2B against zerank-2. We change one setting at a time, from linear to exponential NIST gain or from CAL-A to CAL-B. Each change alone leaves the agreement of RCP-nDCG at worst one pair short in 2019 and complete in 2020. When tied reranker scores share their mean gain, as on ViDoRe v3, 25 pairs are NIST-decisive in 2019 and 41 in 2020. With either TREC-DL judge, RCP-nDCG favors NIST's winner on all of them.

Over the 14 rerankers of the pool panel, the leaderboards of RCP-nDCG and NIST agree weakly in 2019 and moderately in 2020. The Kendall correlations are 0.209 and 0.604, where 1 means identical orders and 0 means unrelated orders. With the four first-stage runs added, the correlations of RCP-nDCG with NIST are 0.477 and 0.725. On DL19, NIST separates none of the ten pairs among the five rerankers that RCP-nDCG ranks highest. The re-assessment grades of Parry et al. (2025) (Appendix H.2) separate none either. No human reference that we use thus orders the five leading rerankers on DL19.

Table 20: Reranker pairs on TREC-DL (14 rerankers, 91 pairs per year). Upper block: pairs that each metric separates by a paired t-test over the year's queries ($p < .05$). Lower block: NIST-decisive pairs on which a metric's mean difference has the sign of NIST's under linear gain. RCP-nDCG uses QWEN3.6-27B and the calibration named in each row (Appendix C.7). Under exponential gain and in the top-150 panel, every disagreement with NIST involves zerank-2.

	DL19		DL20	
	pool	top-150	pool	top-150
Pairs separated				
qrel-nDCG (NIST), linear gain	14	13	37	36
qrel-nDCG (NIST), exponential gain	18	14	36	35
Count-nDCG	56	55	64	64
RCP-nDCG, CAL-A	60	60	70	68
RCP-nDCG, CAL-B	63	61	70	69
Same winner as NIST, linear gain				
RCP-nDCG, CAL-A	14/14	11/13	37/37	36/36
RCP-nDCG, CAL-B	14/14	11/13	37/37	36/36

H.2 NIST POSITIVES THAT RCP SCORES LOWEST

We ask whether the NIST positives that RCP scores lowest point to errors of RCP or of the NIST grades. In the spirit of the label audit of Land & Bikel (2026), we examine the NIST positives (grade 2 or 3) whose calibrated score lies in the bottom quartile of their year's positives. These passages are 1,043 of the 4,167 positives of both years. We compare each with the other positives of its query.

All 14 rerankers rank these passages below the other positives of their query. The within-query AUC is the probability that a reranker ranks another positive of the query above such a passage, where 0.5 means chance. Brackets give 95% CIs from a bootstrap over queries. The mean AUC over the 14 rerankers is 0.879 [0.865, 0.894], and every reranker's interval lies above 0.5. Rerankers still rank them above most passages that NIST grades 0 or 1. Against those passages, the same AUC is 0.295 [0.270, 0.321], below the chance level of 0.5. Stronger rerankers, measured by nDCG@10 on the NIST grades, push the passages further down. Over the 14 rerankers, the Spearman correlation between this nDCG@10 and a reranker's AUC is 0.631 [0.433, 0.776]. Without the zerank models, which learn from LLM preferences, it is 0.955 [0.773, 0.982]. The first-stage retrievers also rank these passages below the other positives, but less consistently (AUC 0.746 [0.719, 0.774]).

On DL19, the re-assessment of Parry et al. (2025) grades all of these passages, blind to the NIST grades. Its re-assessors rate 2.7% of them highly relevant or better, against 40.1% of the other positives. The mean re-assessment grades, on the scale from 0 to 3, are 0.413 for these passages and 1.493 for the other positives. Among NIST grade-3 passages alone, re-assessors rate 3.3% of the 209 low-scored ones highly relevant or better, against 46.7% of the other 488. The 2019 low-gain tail at grade 3 in Figure 4c thus consists largely of passages that the re-assessors also rate low.

The re-assessment also tests the calibrated scores beyond the low-scored positives. Within-query concordance is the share of same-query passage pairs with different grades that a score orders like the grades, where one half is chance. On all passages that both references grade, the concordance of $\tilde{\theta}$ is 0.817 [0.747, 0.860] with NIST and 0.844 [0.819, 0.866] with the re-assessors. The re-assessors' own concordance with NIST is 0.762 [0.702, 0.809]. On DL19, nDCG@10 on the re-assessment grades separates 25 of the 91 reranker pairs. RCP-nDCG@10 picks the winner of each. The rerankers may share preferences with our judge. The quartile cut also selects candidate label errors rather than estimating an error rate.

H.3 HUMAN VERDICTS AND CONTROLS

This subsection gives the full results and controls behind the human verdicts of Section 4.5. Table 21 collects the results of the verdict layer (Appendix D). The main measure and its slices count decided contests (Appendix D.2). Because a metric tie counts only for the other metric, chance for the main measure is not one half. Suppose each verdict were a coin flip. Decided disagreements then favor either metric with probability one half. A contest in which only one metric names a winner

counts, for that metric, only when the verdict names that winner, which happens with probability one half. The study has 156 disagreements with a verdict, 27 contests with a verdict in which only RCP-nDCG names a winner, and 8 in which only qrel-nDCG does. The expected share is then $(156 + 27)/(2 \cdot 156 + 27 + 8) = 52.7\%$. On disagreements alone, chance is one half.

Table 21: Verdict layer. Rates are percentages of decided contests won by RCP-nDCG. Disagreements with a verdict are always decided. Chance is about 53% for the main measure and one half on disagreements. CIs are clustered by query for rates and come from a delete-one-query jackknife for margin AUCs.

Statistic	Est.	95% CI	k/n
Main measure	72.4	[66.0, 78.9]	134/185
NanoBEIR	68.3	[56.4, 80.3]	41/60
BRIGHT	74.6	[64.1, 85.2]	50/67
ViDoRe v3	74.1	[62.8, 85.5]	43/58
without zerank	68.1	[58.7, 77.5]	64/94
one annotator left out (range)	–	70.6–74.4	46 runs
without contests that qrel-nDCG cannot score	70.5	[63.7, 77.4]	122/173
Disagreements only	69.9	[62.7, 77.1]	109/156
without zerank	65.4	[54.9, 75.9]	51/78
one annotator left out (range)	–	68.4–71.8	46 runs
Disagreements, both margins ≥ 0.02	74.8	[67.3, 82.3]	98/131
Disagreements, a margin < 0.02	44.0	[24.1, 63.9]	11/25
Margin AUC: RCP-nDCG	0.810	[0.776, 0.845]	854 picks
qrel-nDCG	0.521	[0.467, 0.576]	854 picks
RCP-nDCG vs. qrel-nDCG, NanoBEIR	0.793 vs. 0.565		284 picks
RCP-nDCG vs. qrel-nDCG, BRIGHT	0.825 vs. 0.501		289 picks
RCP-nDCG vs. qrel-nDCG, ViDoRe v3	0.821 vs. 0.503		281 picks
RCP-nDCG vs. qrel-nDCG, without zerank	0.828 vs. 0.525		402 picks

The main measure's 95% CI lies above chance in each suite and without zerank. A metric built on an LLM judge could favor the three zerank rerankers, which learn from LLM preferences (Pipitone et al., 2025). Without them, the main measure is 68.1% (64 of 94). When we leave out each annotator in turn and recompute the main measure, it stays between 70.6% and 74.4% over 46 runs, inside its CI. In 6 NanoArguAna contests, the displayed documents include the query's own argument, which the NanoBEIR protocol keeps in the pool (Appendix C.5). Annotators graded it 18 times with a mean grade of 0.28 and never Useful or better. This low grade matches the low RCP gain of the argument better than the higher credit that Count-nDCG gives it. Without these contests, the main measure is 72.0% (131 of 182).

Two secondary measures also favor RCP-nDCG. Where a metric names a winner, the verdict agrees with RCP-nDCG in 77.9% (219 of 281) of contests and with qrel-nDCG in 51.9%. Chance is one half. Where both margins of a disagreement are also at least 0.02, verdicts side with RCP-nDCG in 74.8% of contests.

The margin AUC asks whether a metric's margin, its score difference between the two lists, predicts each annotator's pick. One half is chance, and 1 is perfect. The margin AUC never classifies a winner. The tie asymmetry of the main measure therefore does not enter. RCP-nDCG reaches 0.810, against 0.521 for qrel-nDCG, and leads in every suite and without zerank. On disagreements alone, where the two margins point in opposite directions by construction, RCP-nDCG reaches 0.763 and qrel-nDCG 0.346.

If a metric measures what annotators see, its winner should match the verdict more often when its margin is large. Figure 7a bins each metric's absolute margins into five bands. Since bands can differ in how many disagreements they hold, we also bin the margins into fifths over the 254 contests in which both metrics name a winner. Each fifth is reweighted to the study's share of disagreements among all contests with a verdict (54.0%). From the lowest to the highest fifth, the reweighted rate of RCP-nDCG rises by 45.2 pp (95% CI [29.9, 60.4] pp). The reweighted rate of qrel-nDCG rises by 7.7 pp (95% CI [-9.6, 24.9] pp). Both intervals come from a bootstrap over queries.

In the lowest band, below 0.02, the verdict matches the winner of RCP-nDCG in 47.2% of 36 contests (95% CI [30.7, 63.8]%). This share is at chance level. The marginal row of Table 21 instead counts only the decided disagreements in which both metrics name a winner. Where qrel-nDCG names the same list in this band, RCP-nDCG matches 8 of 13 verdicts. Where qrel-nDCG names the other list, RCP-nDCG matches only 8 of 22 verdicts. The margin of qrel-nDCG in these contests is large, with a median of 0.226. Above 0.02, RCP-nDCG matches 82.4% of verdicts. We therefore read differences of RCP-nDCG below 0.02 between two lists of one query as ties.

The controls test for artifacts of the instrument and of the annotators. We replay every verdict with a metric's own gains in place of the annotators' grades. The replayed verdicts match that metric's winner in 98.3% of contests for qrel-nDCG and 87.5% for RCP-nDCG. The instrument thus handicaps RCP-nDCG. A query-blind baseline tests whether verdicts follow overall reranker strength alone. The baseline fits a Bradley–Terry model of reranker strength to the verdicts of all other queries. This model predicts the held-out query's verdicts and matches 65.7% of them. Split pairs meet in several contests and win on some queries and lose on others. Always picking the pair's more frequent winner matches 70.8% of verdicts, whereas RCP-nDCG matches 77.0% and qrel-nDCG 50.4%. Annotators prefer the longer displayed list in 56.7% of contests. We therefore fit a logistic regression that predicts each annotator's pick from both standardized margins and the log ratio of the two lists' lengths. The length weight is -0.02 (95% CI $[-0.28, 0.23]$), and the two margin weights stay at 1.87 for RCP-nDCG and 0.27 for qrel-nDCG. List position cannot drive a verdict, because annotators saw the documents in random order and never saw the lists or their labels. We recompute the verdicts under eight other gain maps. RCP-nDCG leads qrel-nDCG under every map, and the main measure under the study's map lies in the middle of their range. Under the four binary maps, which count relevance from one question upward, the main measure ranges from 64.4% (from Complete) to 75.2% (from Useful).

H.4 LLM JUDGE PANEL AND STUDY 2

This subsection describes the three LLM judges of Study 1 (Appendix F.2) and Study 2 and reports Study 2, which compares the reranker lists that qrel-nDCG and RCP-nDCG prefer.

Judges and verdicts. Studies 1 and 2 use the three judges of Appendix H.5, GLM-5.3-FLASH, DEEPSEEK-4.1-FLASH, and KIMI K3. The three come from model families other than those of our judges. Each judge compares two candidates without knowing their source and never produces a relevance gain. Each case shows a query and two candidates. Except in the controls, one is favored by RCP and the other by the qrels. Every case appears twice, with the labels and the display order swapped. A judge decides a case when both presentations name the same candidate. Under the majority rule, a case counts for a candidate when at least two of the three judges decide it for that candidate. We also report the share among cases that all three judges decide for the same candidate. Under either rule, a share of 50% means no preference. Appendix J.3 prints the prompts.

Study 2. Study 2 asks which of two rerankers' top-10 lists a judge prefers when qrel-nDCG@10 and RCP-nDCG@10 name different winners. On each NanoBEIR dataset, we pair the four best rerankers under qrel-nDCG with the four best under RCP-nDCG. A sign-flip is a query and one of these reranker pairs for which the two metrics name different winners, each by a margin above 0.005. We sampled 800 sign-flips and 200 controls, on which both metrics name the same winner.

By majority, the judges prefer the list of RCP-nDCG in 78.4% of decided sign-flips (514 of 656), with a 95% CI of [74.1, 82.3]% by a query-cluster bootstrap. On the 458 sign-flips that all three judges decide the same way, the share is 82.5%. Each judge alone also prefers this list, with shares from 74.0% to 78.6%. The majority favors it on 12 of the 13 datasets, and HotpotQA is the exception (Appendix F.3). On the controls, the judges pick the list that both metrics prefer in 88.5% of decided cases (154 of 174).

Neither display order nor list length explains this preference. A decided case needs the same verdict in both orders. In single presentations, each judge sides with RCP-nDCG in either display order. GLM-5.3-FLASH, for example, sides with RCP-nDCG in 79.0% of its non-tie verdicts when that list comes first and in 69.3% when it comes second. The list of RCP-nDCG has more characters in 61.5% of sign-flips. The judges pick it in 81.1% of decided cases where it is shorter and in 76.2% where longer.

On BRIGHT, the share for the list of RCP-nDCG is 74.4% over all twelve tasks (467 of 628, 95% CI [70.7, 77.9]%), and 80.2% when all three judges must agree. Unlike Study 1, Study 2 keeps its original lists. On AoPS, LeetCode, and TheoremQA Questions, these lists still contain BRIGHT's excluded documents (Appendix C.5). Both lists show the query's own problem in 175 of the 187 sign-flips on these tasks, and only the list of RCP-nDCG shows it in 9. On these three tasks, the share for the list of RCP-nDCG is 70.3%, against 75.5% on the other nine. On the BRIGHT controls, the judges pick the list that both metrics prefer in 86.3% of decided cases.

Judgments by two of the authors. The authors who judged Study 1 (Appendix F.2) also judged a subset of 80 sign-flips and 20 controls as in Study 1. With their judgments pooled, 79.4% of 126 non-tie judgments on sign-flips prefer the list of RCP-nDCG (95% Wilson CI [71.5, 85.5]%). List length confounds this result. Their judgments favor the list of RCP-nDCG in 89.4% of the non-tie cases where its list is the longer one. Where it is the shorter one, they favor it in 53.3% (95% CI [36.1, 69.8]%), an interval that includes one half. Unlike these judgments by two of the authors, the external human study of Section 4.5 and Appendix D uses paid annotators who are not among the authors.

H.5 External LLM judges

Setup. To test the tournament's order of Count-tied documents (Appendix G.7) directly, we asked LLM judges from model families other than those of our judges to compare two ranked lists. Each case contrasts the RCP side, the list that the tournament's order favors, with the Count side, the list that Count's order favors. The comparison concerns the order within a query and tests neither the calibration across queries nor the leaderboards.

We built four sets of cases, namely per-query disagreements, controls, ideal lists, and minimal pairs, on NanoBEIR, BRIGHT, ViDoRe v3, and TREC-DL, with the paper's 14 rerankers. A per-query disagreement is one query and two rerankers for which Count-nDCG@10 and RCP-nDCG@10 name opposite winners. We sample these cases in equal thirds by the difference in RCP-nDCG@10. Controls are cases in which both metrics name the same winner, ten per third and suite. An ideal-list case sets RCP's ideal top 10 of a query against Count's ideal top 10, whose ties we break at random. A minimal pair sets a reranker's top 10 against the same list in which each group of exactly Count-tied documents takes the tournament's order on its own positions. Count-nDCG@10 scores the two lists alike. Unlike the minimal pairs of Appendix G.7, which reverse the tournament's order, this contrast keeps the reranker's own order as the alternative. The four sets hold 1,147 cases.

The judges are GLM-5.3-FLASH, DEEPSEEK-4.1-FLASH, and KIMI K3, and Appendix J.3 gives the prompt. The label swap, the decision rule, and the majority rule are those of Appendix H.4. Our measure is the share of these decided cases that favor the RCP side, pooled over suites, with 95% bootstrap intervals over queries per suite. For controls, the share counts the list that both metrics prefer. A share of 50% means no preference.

Results. Table 22 gives the shares pooled over suites, and Table 17 the per-suite ideal-list shares. The majority favors the RCP side in all three sets other than the controls. The share is 63.1% on per-query disagreements (95% CI [57.3, 68.5]%), 86.8% on ideal lists, and 84.0% on minimal pairs. Each judge alone also favors the RCP side in every set. Each judge's pooled 95% interval lies above 50% in every set. No suite favors the Count side in any set.

Margin split and restricted sets. Opposite winners can also come from a near tie. We therefore split the per-query disagreements by tie zones. Each metric receives a tie zone, a margin up to which we read a difference between the two lists as a tie. For RCP-nDCG, the zone is 0.02, the margin of RCP-nDCG@5 below which the human verdicts are near chance (Section 4.5). For Count-nDCG, we choose the zone that holds the same share of all per-query reranker pairs as the zone of RCP-nDCG. This share is 23.5% on NanoBEIR, 11.0% on BRIGHT, 22.2% on ViDoRe v3, and 38.6% on TREC-DL. Where both metrics are beyond their zones, the judges side with the winner of RCP-nDCG in 66.1% of decided cases (95% CI [57.1, 75.2]%). Where only Count-nDCG is beyond its zone, they favor neither list, with 52.1% (95% CI [40.0, 63.8]%), as RCP-nDCG itself is inside its tie zone there. The cells of this split are tied to the thirds of the sampling. The split is therefore not independent evidence. Dropping BRIGHT cases that show an excluded document and NanoArguAna cases that show the query's own argument changes the shares little. The majority's shares become 65.2% on

Table 22: External LLM judges, pooled over the four suites: share of decided cases that favor the RCP side (%), with the decided cases in parentheses. The majority column adds 95% query-bootstrap intervals and the number of all cases. For controls, the share counts the list that both metrics prefer. 50% means no preference.

Set	Majority	DeepSeek	GLM	Kimi
Disagreements	63.1 [57.3, 68.5] (317/480)	60.5 (354)	61.3 (297)	66.3 (323)
Controls	75.0 [65.3, 84.3] (76/120)	75.0 (92)	70.9 (79)	77.2 (79)
Ideal lists	86.8 [83.1, 90.3] (333/447)	86.1 (352)	85.3 (306)	86.2 (340)
Minimal pairs	84.0 [73.5, 93.5] (50/100)	76.1 (71)	86.7 (45)	80.8 (52)

per-query disagreements, 87.4% on ideal lists, and 82.2% on minimal pairs, each with a 95% interval above 50%.

I THEORY

This section proves the properties of the gain and of the calibration that Section 3 states. The last subsection relates RCP-nDCG to its rubric-only special case, Count-nDCG. Throughout, all discriminations are positive, $\gamma_c > 0$, and $P_c(\theta) = \sigma\big(\gamma_c(\theta - \beta_c)\big)$.

I.1 THE GAIN, ITS SLOPE, AND ORDER PRESERVATION

IRT measures the precision of the criteria with the test information

$$I(\theta) = \sum\nolimits_c \gamma_c^2 \, P_c(\theta)\big(1 - P_c(\theta)\big), \tag{8}$$

the Fisher information that one yes/no answer to every criterion carries about the ability (Lord, 1980). Higher test information means that the answers locate the ability more precisely.

Lemma I.1 (Properties of the gain)**.** *For any number of criteria $C \geq 1$, discriminations $\gamma_c > 0$, and difficulties β_c, the gain g of Equation (4) has the following properties.*

(i) g is smooth, $0 < g(\theta) < 1$ for every θ, $g(\theta) \to 0$ as $\theta \to -\infty$, and $g(\theta) \to 1$ as $\theta \to \infty$.
(ii) Its slope is $g'(\theta) = I(\theta)/\sum_c \gamma_c > 0$. The gain g is therefore strictly increasing. Under the constraint $\sum_c \gamma_c = C$, the slope is $I(\theta)/C$.

Proof sketch. Each P_c is smooth, lies in $(0, 1)$, and tends to 0 and 1. The gain is a convex combination of the P_c with positive weights, which gives (i). Since $\sigma' = \sigma(1 - \sigma)$, we have $P_c' = \gamma_c P_c(1 - P_c)$. Hence $g' = I/\sum_c \gamma_c$ with every term positive, which gives (ii). □

Proposition 3.1 states that sorting a query's documents by $\tilde{\theta}$, or by any strictly increasing function of it, reproduces their Stage A order.

Proof sketch. The fit sets τ_j to the softplus of an unconstrained parameter. Hence $\tau_j > 0$ (Appendix C.3). Then $x \mapsto f(\tau_j x + \alpha_j)$ is strictly increasing for every strictly increasing f, such as g (Lemma I.1(ii)). Hence $\hat{\theta}_i^{\mathrm{BT}} > \hat{\theta}_{i'}^{\mathrm{BT}}$ holds if and only if $\tilde{\theta}_i > \tilde{\theta}_{i'}$ and if and only if $f(\tilde{\theta}_i) > f(\tilde{\theta}_{i'})$. The ideal ranking sorts the pool by decreasing gain and follows the Stage A order. □

Raw-BT-nDCG applies the gain g to the BT scores centered within each query, without calibration (Appendix B.1). This baseline shares the Stage A order and the ideal ranking of RCP-nDCG and differs only in the gain values (Appendix I.3).

I.2 IDENTIFICATION OF THE CALIBRATION

The BT model depends only on score differences, and the judge compares documents only within a query. Each query's pool is thus its own comparison set. The 2PL likelihood of Equation (2) stays

the same if we stretch the axis and shrink every discrimination by the same factor. The likelihood also stays the same if we shift all offsets and difficulties together. Proposition I.2 shows that these are the only symmetries and that the hard constraints of Section 3.3 remove them.

Proposition I.2 (Identification of the link)**.** *Let every query have two documents with distinct BT scores. Consider two parameter settings* $(\tau, \alpha, \gamma, \beta)$ *and* $(\tau', \alpha', \gamma', \beta')$ *with all scales and discriminations positive.*

- *(i) The two settings give every document the same pass probability on every criterion if and only if, for some* $a > 0$ *and* $b \in \mathbb{R}$, $\tau'_j = a\tau_j$ *and* $\alpha'_j = a\alpha_j + b$ *for every query, and* $\gamma'_c = \gamma_c / a$ *and* $\beta'_c = a\beta_c + b$ *for every criterion.*
- *(ii) Two settings related as in (i) give every document the same gain* $g(\tilde{\theta}_i)$.
- *(iii) Among the settings related as in (i), exactly one satisfies* $\sum_c \gamma_c = C$ *and* $\operatorname{mean}_c \beta_c = 0$.
- *(iv) With the criterion parameters fixed, two settings of* (τ_j, α_j) *give the documents of query* j *the same pass probabilities only if they are equal.*

Proof sketch. The "if" direction of (i) follows by substituting the transformed parameters into Equation (2). For the "only if" direction, equal pass probabilities mean equal logits because the logistic function is injective. Two affine functions of the BT score that agree at two distinct values have equal slopes and intercepts. Part (ii) uses that the gain is a convex combination of the pass probabilities. Part (iii) fixes a by the constraint on $\sum_c \gamma_c$ and b by the centered difficulties. Part (iv) follows from equal logits at two distinct BT scores. □

Under the likelihood alone, other constraints would give the same gains (part (ii)). The hard constraints identify the parameters (part (iii)), and the priors play no part in identification. The constraints only decide where the weak priors on τ_j and α_j act (Appendix C.3).

I.3 The rubric-only special case and the resolution of discrete gains

The Count-nDCG baseline uses Stage B alone. Its gain is a document's pass share (Appendix B.1). Let document i appear in n_i Stage B windows and pass criterion c in S_{ic} of them, and let $T_i = \sum_c \gamma_c S_{ic}$. With the criterion parameters fixed, the rubric-only likelihood of an ability θ is

$$L_i(\theta) = \prod_c P_c(\theta)^{S_{ic}} \big(1 - P_c(\theta)\big)^{n_i - S_{ic}}.$$

Proposition I.3 (Counting is rubric-only IRT scoring)**.** *Fix the criterion parameters. Let* $\hat{\theta}_i$ *maximize* L_i *over the extended real line* $[-\infty, \infty]$. *Then* $g(\hat{\theta}_i) = \mathrm{Count}_i^{\gamma} = T_i / (n_i \sum_c \gamma_c)$, *the discrimination-weighted pass share. With equal discriminations (a Rasch model), the right-hand side is* Count_i, *the gain of Count-nDCG.*

Proof sketch. Since $P'_c = \gamma_c P_c (1 - P_c)$, the log-likelihood derivative is $(\log L_i)'(\theta) = T_i - n_i \sum_c \gamma_c \, g(\theta)$. The strictly concave $\log L_i$ has a finite maximizer exactly when $0 < T_i < n_i \sum_c \gamma_c$, where $(\log L_i)' = 0$ gives the claim. Otherwise the maximum lies at $\pm\infty$, where g equals the pass share. Equal discriminations turn the weighted share into the unweighted one. □

Count-nDCG with weighted shares, raw-BT-nDCG, and RCP-nDCG evaluate the same gain curve g (Proposition I.3). The three metrics differ only in the input to g: the rubric's maximum-likelihood ability, the raw tournament score, and the calibrated score $\tilde{\theta}$. Differences between RCP-nDCG and weighted Count-nDCG therefore come from the tournament and the calibration.

Proposition I.4 (Invisible swaps)**.** *Let a query's pool hold* K *documents whose gains take* L *distinct values, with* n_ℓ *documents at value* ℓ.

- *(i) An exchange of two documents with equal gains leaves DCG@*k *of every ranking unchanged. An exchange of two documents with different gains changes DCG@*k *whenever at least one of them is in the top* k.
- *(ii) The share of document pairs with equal gains is* $s = \sum_\ell n_\ell (n_\ell - 1) / \big(K(K-1)\big) \geq (K - L) / \big(L(K-1)\big)$.
- *(iii) With binary gains,* $s \geq (K-2) / \big(2(K-1)\big)$, *which approaches one half as* K *grows.*

Proof sketch. Part (i) follows because the discount $1/\log_2(r+1)$ strictly decreases in the rank, (ii) follows by Cauchy–Schwarz, and (iii) sets $L = 2$ in (ii). □

Within a query, RCP gains are a strictly increasing function of the BT score (Proposition 3.1). Two documents of a query therefore tie in gain only if they tie in BT score.

Proposition I.5 (Resolution of rubric-only scores)**.** *Let the judge answer the C criteria for every document of a pool of K documents.*

- *(i) If the judge answers each criterion once per document, at most 2^C response patterns occur. If, in addition, every document that passes a criterion also passes all easier criteria, at most $C+1$ patterns occur.*
- *(ii) Any gain that depends only on the response pattern gives equal gains to documents with equal patterns. Under the ordered ladder of (i), at least $\lceil K/(C+1) \rceil$ documents of a pool of K share one pattern.*

Proof sketch. Under the ladder of (i), the passed criteria are the m easiest ones for some $m \in \{0, \ldots, C\}$, and (ii) follows by the pigeonhole principle. □

If the judge answered each criterion once per document, then with $K = 150$ and $C = 5$ the ladder bound would put at least 25 documents on one level. Even the unrestricted $2^5 = 32$ patterns would fall below K. On NanoBEIR, Count-nDCG separates 60.4% of the reranker pairs, close to the 63.2% of RCP-nDCG (Section 4.4). Count-nDCG cannot see how a reranker orders documents with equal pass shares, whereas RCP gains inherit the order of Stage A. Appendix G.7 tests whether the tournament does so correctly.

A document that fails every criterion in every window has no finite rubric-only ability. Count gives all such documents gain 0, whereas RCP orders them by the tournament. Such documents make up 45.8% of TREC-DL pool documents. In the three suites of the human study, such documents have a mean RCP gain of 2.9%. Of the annotators' grades for these documents, 3.2% are Useful or better. Relative to this rate, the gain differs by -0.2 pp for RCP and by -3.2 pp for Count, whose gain is 0. Only the interval of the RCP difference, [-1.9, +1.3] pp, includes 0 (Count: [-5.0, -1.7] pp).

Setting the RCP gain of these documents to 0 changes the paired t-test's decision in 4 of the 182 (dataset, reranker pair) cells on TREC-DL. The change affects 59 of 1,183 cells on NanoBEIR and 12 of 1,092 on BRIGHT, whose lists here still contain each query's excluded documents (Appendix C.5). Setting the gain to 0 reverses no separated pair.

J PROMPTS

This appendix prints the full Stage A and Stage B prompts and summarizes the prompts of the LLM judge studies. At run time, the query replaces `{query_placeholder}`, and the documents of the current window replace `{passages_placeholder}`. We wrote and refined the criteria on 5 queries from each of three NanoBEIR datasets (MS MARCO, FiQA, HotpotQA), reading 20 documents per query. A coding agent edited the wording of the criteria to raise the agreement between the answers of QWEN3.5-397B and the judgments of two of the authors on these documents. These 300 documents make up about 0.3% of the NanoBEIR pools. No grade of our external annotators or of NIST entered this process. BRIGHT, ViDoRe v3, and TREC-DL are fully held out.

J.1 STAGE A PROMPT

The TREC-DL runs of QWEN3.6-27B used exactly this prompt. Together with the shipped rubric, this prompt forms the reference cell of the factorial in Appendix G.6. The stored runs of NanoBEIR, BRIGHT, and ViDoRe v3 do not record their Stage A prompt file. We therefore do not claim that this text produced their scores unchanged. The prompt describes the score range as a logit scale, which the soft preferences of Section 3.1 read as confidence gaps.

Stage A prompt (tournament)

You are an **expert relevance ranker** for an information-retrieval system. Your job is to reorder a given list of documents so that the most useful ones for answering the user's search query appear first, and to assign a calibrated relevance score to each document.

Input

- **Query**: a natural-language question or search statement.
- **Documents**: entries labelled `doc_1`, `doc_2`, ... `doc_{num_documents_placeholder}`, presented in arbitrary order.

How to decide relevance

1. **Understand the query**
   - Identify the main topic, entities, intent, and any required answer type (factual, navigational, how-to, comparison, causal, opinion, transactional, list/aggregation, etc.).
   - If the query is **genuinely ambiguous** (e.g., "apple" could mean the fruit or the company), commit to the **single most probable interpretation** given the documents and note it in your reasoning. Score every document against that one interpretation – do not mix interpretations across documents.
   - Surface **explicit constraints** (e.g., "free", "Python", "2024", "without X") and **implicit constraints** (e.g., language, recency, platform) that a relevant document must satisfy. Documents that violate a hard constraint should be scored substantially lower than those that satisfy it, even when the violating document has strong topical overlap.
   - Expand the query with obvious synonyms or related terms (e.g., "clear" $\rightarrow$ "clearing services", "condensate pump" $\rightarrow$ "dehumidifier pump").

2. **Score each document** (criteria listed in order of importance)
   - **Answer directness** – sentences that directly answer the query are the strongest signal.
   - **Specificity** – mentioning the exact entity or detail asked for (a product name, a service, a concrete fact) outranks vague or generic mentions.
   - **Semantic relevance** – discussing the same concept even without exact keyword overlap deserves a high score.
   - **Comprehensiveness** – addressing multiple facets of a multi-part query is better than addressing only one.
   - **Constraint satisfaction** – documents that meet both explicit and implicit query constraints rank higher; those that violate a constraint are penalised proportionally to its importance.
   - **Keyword presence** – exact query terms (or synonyms) in context are a supporting signal, not a primary one.
   - **Accuracy & credibility** – correct information from a trustworthy source.
   - **Penalize irrelevance** – unrelated topics, wrong entities, or generic filler should receive negative scores. Truncated snippets: judge only visible content. Near-duplicates: rank the more complete version higher. Boilerplate or login walls: score ≤ -4.

Bias checks (actively counteract these):

- The document order is arbitrary – your ranking must not be influenced by it.
- Longer documents are not inherently more relevant – judge by information density.
- **Context independence** – score each document against the query on the absolute scale above, regardless of what other documents appear in this batch. Anchor by imagining a hypothetical document that perfectly answers the query ($+5$) and one that is completely off-topic (-5); place each real document on that fixed ruler. A document's score must not change because a stronger or weaker document is present.

3. **Relative ordering**
   - All high-relevance documents must appear before any low-relevance ones.

Output format

Return **only** a JSON object with three fields – no other text.

`"reasoning"`: a brief string summarising how you interpreted the query and the key factors that determined your top-ranked and bottom-ranked documents. This is your self-check – use it to verify that your ranking and scores are internally consistent before finalising.

`"ranking"`: document identifiers (`doc_1`, `doc_2`, ...) from most to least relevant. The array must be sorted by strictly descending score.

`"scores"`: for **every** document, a calibrated relevance score on a logit scale from -5.0 (completely irrelevant) to $+5.0$ (perfect answer). Differences on this scale should reflect your confidence:

- A 1-point gap $\approx 73\,\%$ sure the higher doc is better (moderate)
- A 2-point gap $\approx 88\,\%$ sure (strong)
- A 3-point gap $\approx 95\,\%$ sure (overwhelming)
- A 0-point gap $\approx 50\,\%$ – a coin flip

Score anchors:

$+4$ to $+5$: directly and completely answers the query; could be the sole result
$+2$ to $+4$: highly relevant, key facts present, may be incomplete
0 to $+2$: partially relevant, right topic, doesn't clearly answer
-2 to 0: tangentially related, shares keyword/domain, mostly off-topic
-5 to -2: irrelevant, wrong topic/entity, or noise

Use tenths freely (e.g., 2.3, -0.7). **Always assign distinct scores.** The **size of each gap** must match how confident you are that the higher-scored document is truly better (see the gap-confidence table above). When documents are nearly indistinguishable, keep gaps small (0.1–0.2 $\approx$ a coin flip); use gaps of ≥ 1.0 only when you are genuinely confident. Never give every document the same score.

Example for 4 documents:

```
{
  "reasoning": "The query asks for X. doc_3 directly answers with a specific
      fact; doc_2 is about an unrelated topic.",
  "ranking": ["doc_3", "doc_1", "doc_4", "doc_2"],
  "scores": {"doc_3": 3.5, "doc_1": 1.5, "doc_4": 0.5, "doc_2": -2.5}
}
```

Recap

1. Read the query – extract core concepts, intent type, and both explicit and implicit constraints.
2. For each document, judge how well it addresses those concepts and satisfies those constraints.
3. Assign a calibrated logit-scale score to every document.
4. Sort document identifiers from highest to lowest score.
5. Write a brief reasoning string to verify your ranking is consistent.
6. Return the JSON object and nothing else.

Query

```
{query_placeholder}
```

Documents

```
{passages_placeholder}
```

J.2 Stage B prompt

The Stage B prompt asks five yes/no questions about each document of a window. The judge received exactly this prompt on ViDoRe v3 and TREC-DL. The NanoBEIR and BRIGHT runs used an earlier version of the prompt, with the same wording of the criteria C1 to C5. We therefore do not claim that the printed text produced their scores unchanged.

Stage B prompt (rubric)

Role

You are an **expert relevance assessor** for an information-retrieval system. For each document, answer five binary diagnostic questions that measure how well the document serves the user's query.

Instructions

1. **Understand the query**
   - Identify the core information need, intent type, and any required answer type (factual, navigational, how-to, comparison, causal, opinion, transactional, list/aggregation, etc.).
   - If the query is **genuinely ambiguous**, commit to the **single most probable interpretation** and evaluate every document against it consistently.
   - Surface **explicit constraints** (e.g., "free", "Python", "2024", "without X") and **implicit constraints** (e.g., language, recency, platform).
   - For **argumentative or stance queries** (e.g., "Should X be allowed?", claims to be verified), a document is relevant (C1) if it discusses the same debate topic – even from the opposite side. Information utility (C2) includes evidence or reasoning that bears on the claim, whether supporting or opposing.
2. **Evaluate each document independently** against the five criteria below. Each criterion is a standalone yes/no question – answer 1 (yes) or 0 (no). Criteria are **not hierarchically dependent**: a document may pass any criterion regardless of whether it passes the others.
3. **Coherence review** – after scoring all criteria, briefly verify that your ratings are internally consistent. If you notice any apparent contradiction, re-read the relevant criterion definition and the document before finalizing.

Bias checks – actively counteract these:

- Document order is arbitrary – do not let it influence your judgments.
- Longer documents are not inherently more relevant – judge by substance, not length.
- Evaluate each document against the query on an absolute standard, regardless of what other documents appear in this batch.
- C5 is a **high bar** – most documents should fail it. Do not let a document's length, apparent effort, or broad coverage bias you toward awarding it.

Criteria

C1 – Topical Relevance

"Does this document address a topic that is clearly related to the query's information need – not just loosely related or in the same broad domain?"

Award 1 if a two-sentence summary of the document would mention the same topic as the query – i.e., the connection is immediate and obvious. Award 0 only if the connection requires reasoning through shared methods, broader categories, or domain overlap.

C2 – Information Utility

"Does this document contain specific, useful information – facts, data, procedures, or explanations – that would help address the query?"

Award 1 if a user could extract at least one concrete fact, data point, procedure step, or example that directly bears on the query. Award 0 if the document only provides general background, definitions, or context without specific applicable information.

C3 – Entity/Detail Match

"Does this document specifically name and discuss the particular entity, product, concept, or detail mentioned in the query – not merely a related alternative or a broader category that includes it?"

Award 1 if the exact entity or detail from the query appears by name and the document provides at least some specific information about it – not just a mention in a list, heading, or aside. Award 0 if the document covers only a broader category, a competing or related entity, or never names the specific entity.

C4 – Direct Answer

"Does this document explicitly and directly answer or resolve the query's primary question or information need?"

Award 1 if the document contains a clear, direct response to the main question – even if it doesn't cover every secondary aspect. Award 0 if the user would still need to search elsewhere for the core answer.

C5 – Thorough Treatment

"Does this document treat the query's topic in depth – providing detailed explanations, thorough coverage, or comprehensive detail rather than a superficial or cursory mention?"

Award 1 only if the document provides genuinely deep, substantive coverage that goes well beyond a basic answer – e.g., explaining mechanisms, providing supporting evidence, addressing edge cases, or offering practical detail that a knowledgeable reader would expect. Award 0 if the treatment is surface-level, merely lists aspects without elaboration, restates the same point in different ways, or is long without adding real depth. A lengthy or broad document is not automatically thorough.

Query

`{query_placeholder}`

Documents

`{passages_placeholder}`

Output Format

Return a JSON object. For each document, provide:

- `"reasoning"`: a brief one-line justification
- `"criteria"`: an object mapping C1–C5 to 1 or 0

Example:

```
{
  "doc_1": {
    "reasoning": "Official Python 3.11 docs on sorted(); exact match,
          comprehensive, authoritative.",
    "criteria": {"C1":1,"C2":1,"C3":1,"C4":1,"C5":1}
  },
  "doc_2": {
    "reasoning": "Python sorting tutorial for 3.9. Answers the question but
          explanations stay surface-level and omit edge cases.",
    "criteria": {"C1":1,"C2":1,"C3":1,"C4":1,"C5":0}
  },
```

```
  "doc_3": {
    "reasoning": "Sorting algorithms theory. Relevant topic but no Python code
          or practical answer.",
    "criteria": {"C1":1,"C2":0,"C3":0,"C4":0,"C5":0}
  },
  "doc_4": {
    "reasoning": "Java sorting tutorial. Sorting concepts transfer, but does
          not cover Python's sorted().",
    "criteria": {"C1":1,"C2":1,"C3":0,"C4":0,"C5":0}
  },
  "doc_5": {
    "reasoning": "Unrelated article about wildlife conservation.",
    "criteria": {"C1":0,"C2":0,"C3":0,"C4":0,"C5":0}
  }
}
```

J.3 META-JUDGE AND EXTERNAL-JUDGE PROMPTS

Studies 1 and 2 send their prompts to the three judges of Appendix H.4, each at its maximum reasoning effort, with no output cap and the provider's default temperature. The two system prompts are those of the original study scripts. The judges answer in the shared response format printed below. Each prompt also shows the dataset's MTEB retrieval-task instruction. Study 1 asks which of two candidate sets of relevant documents better serves as the query's annotation, weighing relevance, precision, and coverage. Study 2 compares two rerankers' top-10 lists, adds ranking quality to these criteria, and shortens each document to 1,500 characters. Two of the authors judged part of Study 2 and saw the same shortened lists (Appendix H.4).

Shared response format

Respond with exactly the following three fields, in this order:

JUDGMENT: A concise explanation (1–3 sentences) naming a concrete, content-level reason for your choice.
CONFIDENCE: Exactly one of `"high"`, `"medium"`, or `"low"`.
VERDICT: Exactly one of `"A"`, `"B"`, or `"Tie"`.

In the comparison of Appendix H.5, each case also shows a task description of what counts as relevant in its dataset. The NanoBEIR descriptions ask, for example, for documents that refute a claim or for duplicate questions. BRIGHT asks for passages that answer a question in the task's discipline. TREC-DL cases use the MSMARCO web-search description, and ViDoRe v3 cases ask for the pages with the needed information.

Appendix G.7 and Appendix H.5 use the same three judges, GLM-5.3-FLASH, DEEPSEEK-4.1-FLASH, and KIMI K3, with a prompt of their own. This prompt asks which of two ranked lists better serves the query and weighs top ranks more heavily than lower ones. The prompt tells the judges to decide by content match, to treat near-duplicates as equivalent, and to answer Tie only when neither list is defensibly better. Each response is a JSON object with the fields `choice` (A, B, or Tie), `confidence`, and a one-sentence `reason`. The user prompt shows the task description, the query, and the two lists without document identifiers or metric hints. A document longer than 2,000 characters is cut at its last sentence boundary before this limit, and the marker `[...]` shows the cut. Each case is sent twice, with the labels and the display order of the two lists swapped. The full text of every prompt in this appendix is part of the release (Appendix K).

K ASSETS AND RELEASE

Table 23 lists the datasets with their sources and licenses.

Models and software. Our three judges are open-weight models under Apache 2.0: the NVFP4 release of QWEN3.5-397B, QWEN3.6-27B, and GPT-OSS-120B. We serve them with SGLang. Of the three first-stage retrievers, Octen-Embedding-8B is also under Apache 2.0, and BM25 runs in the MIT-licensed `bm25s`. We used the three panel judges (Appendix H.4), GLM-5.2 (Appendix G.4),

Table 23: Datasets and third-party relevance data, with each provider's stated license or terms of use.

Asset	Source and use	License or terms
NanoBEIR	Nano versions of 13 BEIR datasets in MTEB (Thakur et al., 2021; Muennighoff et al., 2023; Zeta Alpha, 2024), `mteb/Nano*Retrieval`; 649 queries	CC BY 4.0 on the dataset cards; the BEIR source datasets state their own terms
BRIGHT	12 tasks (Su et al., 2025), `mteb/BrightRetrieval`; 1,366 evaluated queries	CC BY 4.0
ViDoRe v3	8 corpora (Loison et al., 2026), `vidore/vidore_v3_*`; 2,419 questions in their original language, judged as OCR text	CC BY 4.0
TREC-DL	MS MARCO passages (Nguyen et al., 2016) with the NIST grades of the 2019 and 2020 tracks (Craswell et al., 2020; 2021); 43 and 54 queries	MS MARCO terms, non-commercial research only; NIST grades public from TREC
DL19 re-assessments	Parry et al. (2025), public repository at commit `12969fb`; analysis only, not redistributed	Paper CC BY 4.0; no license file in the repository

Cohere Embed v4, and Voyage AI's rerankers under their providers' terms. Table 3 gives the licenses of the rerankers.

Human annotations. The human grades of Sections 4.2, 4.3, and 4.5 come from an annotation platform (Appendix D.3). The platform stored the annotators' names and email addresses. In the released data, one-way hashes replace them, and no table maps the hashes back. We have not screened the annotators' free-text comments for personal information beyond replacing email addresses. We therefore do not release these comments.

Release. The supplementary material holds the code of both stages, the calibration, the metric, and the full text of every judge prompt of Appendix J. We will release this code and these prompts publicly. The supplementary material also holds the calibrated scores of every judged query-document pair, the parameters that turn them into gains, and the 7,080 human grades on 2,268 documents. The BRIGHT labels also cover the documents that BRIGHT excludes for some queries (Appendix C.5). Our rankings and tables remove these documents, except in the few follow-up analyses that say so.

L Extended Related Work

We compare RCP in more detail with three lines of work that the main text covers only briefly.

Relative and absolute judgments together. Yan et al. (2024) observe that an LLM ranks well from pairwise prompts, whereas its pointwise relevance labels rank documents less well and its pairwise scores carry no absolute meaning. They also find that pairwise scores are not calibrated across queries. They perturb each document's pointwise label as little as possible until the labels respect the pairwise order. RCP instead keeps the tournament order fixed and fits only a scale τ_j and an offset α_j per query. Binary criteria shared by all queries set these two parameters, as shared anchor items place two test forms on one scale in IRT linking (Lord, 1980). Similarly, label models infer true labels from noisy annotators by estimating each annotator's expertise and each item's difficulty (Whitehill et al., 2009).

Tournament gains and per-query scores. Concurrently, SABER-Math builds labels for mathematical retrieval from per-query BT tournaments and rescales each query's scores to a fixed range before computing nDCG (Georgiev et al., 2026). The zELO method turns LLM pairwise preferences into per-query Elo scores and centers each query's scores at zero to train the zerank rerankers (Pipitone et al., 2025). This centering fixes each query's zero point, but it makes neither the zero points nor the units comparable across queries. In RCP, the rubric shared by all queries supplies both.

Rubrics and calibrated LLM judges. Rubric and checklist judges ask an LLM about several criteria and aggregate the answers, for example by counting, summing, or taking the best grade (Farzi & Dietz, 2024; 2025; Lee et al., 2025; Ye et al., 2024). Up to the criterion weights, Count-nDCG (Equation (5)) is the retrieval analogue of these methods and the rubric-only special case of RCP

(Proposition I.3). Other work calibrates LLM judges against human labels (Liu et al., 2024; Hashemi et al., 2024; Takehi et al., 2025). RCP uses no human labels for calibration. Its rubric only sets the scale and offset of each query's tournament scores.