AI and Data Science Journal
Online ISSN: 3078-7394
DOI: https://doi.org/10.61784/adsj4043

UNCERTAINTY-AWARE LLM-AS-A-JUDGE EVALUATION FOR RELIABLE SEARCH AND RECOMMENDATION SYSTEMS: CALIBRATED CONFIDENCE AND IMPACT-AWARE SELECTIVE HUMAN REVIEW

GaiKe Wang[1*], PengFei Li[2], EnMiao Feng[2], TianJun Mo[2]
[1]*New York University, New York, NY 10012, USA.*
[2]*Duke University, Durham NC 27708, USA.*
[*]*Corresponding Author: GaiKe Wang*

Abstract: Large language models (LLMs) are increasingly used as judges that produce relevance labels for evaluating search and recommendation systems, yet their graded judgments fluctuate with the prompt, the model and the assessment criterion, disagree with human assessors in a topic-dependent way, and come without trustworthy confidence. This paper presents CARE (Consistency-Aware calibration with Risk- and Evaluation-impact-driven review), an uncertainty-aware framework that turns a panel of LLM judgments into calibrated relevance distributions and spends a limited human budget where it matters most for evaluation outcomes. CARE encodes multi-dimensional consistency signals across repeated runs, prompts, LLMs, criteria and fine-tuning; calibrates them with a hierarchical multinomial model whose topic-specific grade biases receive an empirical-Bayes prior and a Laplace posterior; and routes pairs to assessors by expected ordinal error weighted by their impact on system comparisons, complemented by online posterior updates, soft expected-gain qrels and risk-targeted auto-acceptance. On the TREC DL 2023-based LLMJudge benchmark with 33 real LLM label sets (GPT-4o, Llama-3-8B/70B and FLAN-T5 judges using zero-shot, chain-of-thought, few-shot, criteria-decomposed and fine-tuned pipelines), changing only the prompt flips 22–46% of graded labels, whereas repeated sampling flips fewer than 0.3%. With ten human labels per topic, CARE raises Cohen's κ with TREC assessors from 0.273 (majority vote) to 0.412 and reduces the expected calibration error from 0.150 to 0.036. Reviewing 10% of the pairs selected by CARE lifts the Kendall's τ between system rankings from 0.642 (random review) to 0.895, and $\tau \geq 0.85$ is reached with 7.7% of the judging effort.
Keywords: LLM-as-a-judge; Relevance assessment; Uncertainty calibration; Selective human review; Test collections; Search and recommendation evaluation

1 INTRODUCTION

Offline evaluation of search engines and recommender systems rests on relevance labels, i.e., graded statements of how well an item satisfies an information need. Collecting these labels from trained assessors is slow and expensive, which has motivated the use of large language models (LLMs) as relevance assessors [1-4]. Deployments at web scale [1], open reproductions such as UMBRELA [4], industrial studies in product search and LLM evaluators for recommendation explanations all report encouraging agreement with humans [5-6], and the LLMJudge challenge at SIGIR 2024 collected dozens of LLM label sets on topics of the TREC 2023 Deep Learning track [7-9].

Three properties of LLM judges nevertheless limit their reliability. First, judgments are sensitive to seemingly innocuous design choices: paraphrased prompts change label quality unpredictably [1], formatting changes alter accuracy [10], and candidate order biases pairwise verdicts [11]. Second, disagreement with human assessors is systematic rather than random; it varies across topics and grades and can be exploited adversarially [12-13], which has led to the position that LLMs should not replace human assessment [14]. Third, LLM judges rarely expose calibrated confidence: the fraction of agreeing votes in a judge panel looks like a probability but, as shown in Section 6, it is strongly overconfident. Without reliable confidence it is unclear which labels can be trusted and which should be checked by a person, so the human effort that remains in any practical pipeline [2] is spent blindly.

Ensembling several judges and selecting the strongest prompt reduce [1, 15-16], but do not remove, these issues. Selective evaluation with agreement guarantees has been developed for pairwise preference judgments [17], and prediction-powered inference corrects aggregate estimates but does not decide which items deserve a human label [18-19]. What is missing for search and recommendation evaluation is an end-to-end mechanism that treats inconsistency across prompts, models and criteria as an uncertainty signal instead of noise to be averaged away, calibrates graded relevance under topic shift, and allocates a limited human budget so that the final test collection ranks systems as human assessors would.

This paper proposes CARE, an uncertainty-aware evaluation framework built around these requirements, and makes four contributions. (1) A multi-dimensional consistency analysis of 33 real LLM label sets that separates sampling stochasticity, prompt, model, criterion and topic effects; prompt choice, not sampling, dominates inconsistency, and LLM leniency relative to humans varies by more than 1.5 grade points across topics. (2) A hierarchical topic-adaptive calibrator that maps compact multi-dimensional consistency (MDC) features to calibrated distributions over relevance

© By the Author(s) 2024, under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0).

grades, with an empirical-Bayes prior over topic-specific grade biases and Bayesian predictive integration. (3) An evaluation-impact-aware selective review policy that ranks pairs by expected ordinal error multiplied by their influence on system comparisons, combined with online posterior updates, soft expected-gain qrels and risk-targeted auto-acceptance. (4) An evaluation at label, topic and system levels against eleven aggregation and review baselines, with ablations and five judge-panel configurations.

With ten human labels per new topic, CARE improves Cohen's κ against TREC assessors from 0.273 for majority voting to 0.412 and lowers the expected calibration error (ECE) from 0.150 to 0.036. Reviewing 10% of the pairs selected by CARE yields a Kendall's τ of 0.895 between the system ranking induced by the final qrels and the human ranking, compared with 0.642 for random review.

2 RELATED WORK

2.1 LLMs as Relevance Assessors

A single labeled example can already guide a model to fill relevance holes [3], and Faggioli et al. [2] outlined a spectrum of human–LLM collaboration for relevance judgment. Thomas et al. [1] selected an LLM and prompt against first-party searcher feedback and found LLM labels as accurate as those of third-party assessors, while observing unpredictable sensitivity to paraphrases. UMBRELA reproduced this assessor with open tooling [4], and Mehrdad et al. fine-tuned LLMs for query–item relevance in product search [5]. The LLMJudge challenge benchmarked prompting [7], fine-tuning and ensembling strategies; rubric- and criteria-based judges [20], fine-tuned judges originally built for query performance prediction and ensembles of small judges were among the explored designs [15, 21]. Critical studies show that LLM assessors can be manipulated by injected query terms [12], that runs can be crafted to exploit them [13], and that using them for system comparison raises circularity concerns [14]. In recommendation, LLM evaluators have been shown to rate explanation quality along several user-perceived aspects [6].

2.2 Reliability and Uncertainty of LLM Judges

LLM-as-a-judge was popularized for evaluating chat assistants [22]. Later work documented positional bias [11], prompt-format sensitivity and intra-model bias [10], motivating panels of diverse judges [16]. Confidence elicitation studies report that verbalized and sampling-based confidences of LLMs are poorly calibrated [23]. Jung et al. proposed selective evaluation in which a judge abstains or escalates when its confidence is low [17], with a guarantee on agreement with humans for pairwise preferences. CARE shares the selective view but targets graded relevance, models topic-level shift explicitly and optimizes the selection for the downstream system ranking.

2.3 Calibration and Label Aggregation

Temperature scaling is the standard post-hoc calibrator for neural classifiers [24]. Aggregating noisy annotators has a long history: Dawid and Skene estimate annotator confusion matrices with EM [25], and Bayesian classifier combination places priors on them [26]. Label-distribution shift between training and deployment can be corrected by re-estimating class priors [27]. CARE combines these ideas in a supervised hierarchical form: topic-specific grade biases act as shifted priors whose distribution is learned from historical topics.

2.4 Efficient and Human-in-the-Loop Evaluation

Minimal test collections select documents whose labels discriminate between systems [28], and AP correlation emphasizes agreement at the top of a system ranking [29]. Selective classification trades coverage for risk [30], and prediction-powered inference [18], used by ARES for retrieval-augmented generation [19], debiases aggregate metrics with a small labeled set. CARE connects these lines by choosing human labels with calibrated, impact-weighted uncertainty and by producing per-pair soft labels for metric computation.

3 PROBLEM FORMULATION

Let Q be a set of topics and D_q the judged pool of topic q. For each pair $i=(q,d)$ a panel of J LLM judges returns graded labels $\ell_{ij} \in \{0,1,2,3\}$, and a human assessor would return y_i. A set S of retrieval or recommendation systems is evaluated with nDCG@10 computed from a qrel vector g. A practitioner has human labels for historical topics and a budget of B new human labels for the current topics. The goal is to construct, for every pair, either a human label or an automatic surrogate such that the resulting evaluation agrees with full human judging at three levels: the label level (Cohen's κ and quadratic-weighted κ), the topic level (correlation of per-topic nDCG@10) and the system level (Kendall's τ and τ_{AP} between system rankings) [29]. A second requirement is a calibrated predictive distribution $p(y|x_i,q)$, so that confidence can drive both human routing and automatic acceptance.

4 THE CARE FRAMEWORK

Figure 1 gives an overview. CARE first encodes how the judge panel agrees with itself, then calibrates this evidence with a hierarchical topic-adaptive model, and finally uses the calibrated distribution to route pairs to assessors, to accept confident labels and to compute soft qrels.

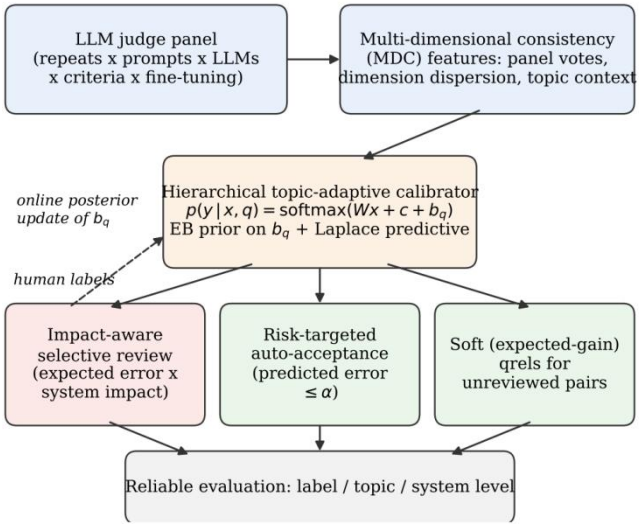


Figure 1 Overview of CARE

Panel judgments are summarized by multi-dimensional consistency features, calibrated per topic, and used for impact-aware human review, risk-targeted acceptance and soft-label evaluation; human labels update the topic posterior online.

4.1 Multi-Dimensional Consistency Encoding

Instead of feeding raw judge labels to a calibrator, CARE summarizes agreement within the panel along interpretable dimensions. Panel-level features are the vote histogram, the mean and standard deviation of the labels, the vote entropy, the majority fraction and the pairwise disagreement rate. For a group G of judges that differ in a single factor, the pairwise disagreement of pair i is

$$\delta_G(i) = 2 / (|G|(|G|-1)) \cdot \Sigma_{j<j' \in G} \, 1[\ell_{ij} \neq \ell_{ij'}]. \quad (1)$$

Groups are formed for the repeat dimension (same LLM and prompt, independent runs), the prompt dimension (same LLM, different prompts) and the model dimension (same pipeline, different LLMs); for each group CARE records δ_G, the standard deviation, the range and the mean label. The criterion dimension is captured by the signed gap between the mean labels of holistic and of criteria- or rubric-decomposed judges, and the adaptation dimension by the gap between prompted and fine-tuned judges. Three label-free topic-context features (the topic mean of panel labels, the topic mean vote entropy and the within-topic percentile of the pair's panel mean) describe how lenient and how contested the panel is on a topic. No feature requires human labels; for the 25-judge panel the MDC vector has 30 dimensions.

4.2 Hierarchical Topic-Adaptive Calibration

Let $\varphi(x_i)$ denote the standardized MDC vector. CARE models

$$p(y_i=k|x_i,q) = \mathrm{softmax}_k(w_k^\top \varphi(x_i)+c_k+b_{q,k}), \quad (2)$$

where W and c are shared across topics and $b_q \in \mathbb{R}^4$ is a topic-specific grade-bias vector. The bias captures that LLMs are more lenient than humans on some topics and stricter on others, and that the usage of individual grades shifts rather than the whole scale (Section 6.2 compares it with a scalar ordinal offset). A Gaussian prior $b_q \sim \mathrm{N}(0, (2\lambda)^{-1}\mathrm{I})$ is placed on the biases. The shared parameters are fitted by L2-regularized multinomial logistic regression on historical topics, and λ is set by empirical Bayes from the dispersion of per-topic bias estimates on those topics:

$$\hat{\lambda} = [\, 2 \cdot \mathrm{mean}_{q,k} \, (\tilde{b}_{q,k}-\bar{b}_q)^2 \,]^{-1}, \quad (3)$$

where $\tilde{b}_q$ is a weakly regularized MAP estimate on historical topic q and $\bar{b}_q$ its mean (softmax is invariant to a common shift). For a new topic with a set R_q of reviewed pairs, the posterior of b_q is approximated around its MAP estimate (Laplace approximation):

$$\hat{b}_q = \mathrm{argmax}_b \, \Sigma_{i \in Rq} \log p(y_i|x_i,q;b) - \lambda \|b\|^2, \quad (4)$$

with Laplace covariance $\Sigma_q = (\Sigma_{i \in Rq}(\mathrm{diag}(p_i)-p_i p_i^\top) + 2\lambda \mathrm{I})^{-1}$, where p_i is the grade distribution of pair i at $\hat{b}_q$. The predictive distribution integrates over this posterior with M=64 Monte-Carlo samples $\varepsilon_m \sim \mathrm{N}(0, \mathrm{I})$:

$$\bar{p}(y|x,q) \approx (1/M) \, \Sigma_m \, \mathrm{softmax}(W\varphi(x)+c+\hat{b}_q+\Sigma_q^{1/2}\varepsilon_m). \quad (5)$$

When R_q is empty the prior is used, so cold-start topics receive appropriately broadened distributions instead of overconfident ones.

4.3 Evaluation-Impact-Aware Selective Review

Given $\bar{p}$, CARE measures how wrong the automatic label is expected to be on the ordinal scale:

$$e(i) = \Sigma_k \bar{p_k}(i)\,|k-\mu(i)|, \qquad \mu(i) = \Sigma_k\, k\, \bar{p_k}(i). \tag{6}$$

Not every label error matters equally for evaluation. Following the intuition of minimal test collections [28], a label that enters the top ranks of some systems but not of others changes their relative nDCG@10. For a system s that ranks pair i at position $r_s(i)$, let $c_s(i) = 1[r_s(i)\leq 10] \,/\, (\log_2(1+r_s(i)) \cdot \mathrm{IDCG}_q)$, where IDCG_q is approximated with label-free panel means. The impact of pair i and its review priority are

$$\omega(i) = \mathrm{std}_{s\in S}\, c_s(i) + 0.1 \cdot \mathrm{mean}_{s\in S}\, c_s(i), \tag{7}$$

$$\pi_\lambda(i) = e(i) \cdot [(1-\lambda) + \lambda\omega(i) \,/\, \bar{\omega}], \tag{8}$$

where $\bar{\omega}$ is the mean impact of the candidates. λ=1 gives the impact-aware policy used in the main experiments, λ=0 gives expected-error sampling, and intermediate values trade label-level for system-level agreement (Section 6.3). Algorithm 1 summarizes the procedure. A few random seed labels per topic are collected first so that each topic bias is estimated from an unbiased sample; the remaining budget is spent in R rounds, after each of which the topic posteriors are updated. Because selection depends only on covariates, the conditional likelihood in (4) remains valid for actively chosen labels.

Algorithm 1 CARE Selective Review

Input: panel labels; systems S; calibrator $(W, c, \hat{\lambda})$ fitted on historical topics; budget B; seed size m_0; rounds R; trade-off λ.

1: for each new topic q: send m_0 random pairs to assessors, $R_q \leftarrow$ seeds
2: for $r = 1, \ldots, R$ do
3: update $(\hat{b_q}, \Sigma_q)$ from R_q for every topic (Eq. 4)
4: compute $\bar{p}$ (Eq. 5) and π_λ (Eq. 8) for unreviewed pairs
5: send the $\lceil B_{left}/(R-r+1) \rceil$ top-priority pairs to assessors
6: end for
7: qrels: human labels where reviewed, soft gains (Eq. 9) elsewhere
8: optionally auto-accept A_α (Eq. 10) and route the rest to assessors

4.4 Soft Qrels and Risk-Targeted Acceptance

For unreviewed pairs CARE uses the expected gain $\mu(i)$ of (6) as a soft label when computing nDCG@10:

$$g(i) = y_i \text{ if } i \text{ is reviewed, and } g(i) = \mu(i) \text{ otherwise.} \tag{9}$$

Soft gains avoid the rounding bias of hard labels and propagate calibrated uncertainty into the metric. Calibrated confidence also supports an acceptance rule: sorting pairs by confidence $\max_k \bar{p_k}(i)$, CARE accepts the largest prefix A whose predicted error does not exceed a target α,

$$A_\alpha = \mathrm{argmax}_A\, |A| \quad \text{s.t.} \quad (1/|A|)\, \Sigma_{i\in A}\, (1 - \max_k \bar{p_k}(i)) \leq \alpha, \tag{10}$$

and routes the remaining pairs to assessors. The rule requires no additional labels; its reliability rests entirely on calibration, which Section 6.4 examines.

5 EXPERIMENTAL SETUP

5.1 Data and Judge Panels

Experiments use the LLMJudge test collection [7], built from the passage ranking task of the TREC 2023 Deep Learning track [9], whose topics include both real and LLM-generated queries. It contains 25 topics and 4,423 query–passage pairs judged by NIST assessors on a four-point scale (2,005 irrelevant, 1,233 related, 808 highly relevant and 377 perfectly relevant). We use the 33 LLM label sets submitted to the challenge by eight teams in 2024; per-judge Cohen's κ against the human labels reproduces the published values exactly (e.g., 0.2863 for the strongest GPT-4o judge). Table 1 summarizes the judges. Eight supervised or fine-tuned pipelines are reserved as reranking signals of the evaluated systems and never enter a judge panel, which prevents direct circularity between rankers and judges [2, 14]. The remaining 25 judges form the full panel; Section 6.5 also studies four smaller panels.

Table 1 Judge Panel (25 LLM Judges) and Reserved Reranking Signals

Group	#	LLM	Prompting / pipeline
GPT-4o prompted	7	GPT-4o	zero-shot, zero-shot + reasoning, few-shot, UMBRELA numeric / semantic / min-ensemble, binary filter + 3-score average
Llama-3 prompted	14	Llama-3-8B / 70B	zero-shot (3 runs), CoT (3 runs), filter + average (8B, 70B), 4-criteria decomposition (prompt, sum, naive Bayes), criteria CoT, query-generation matching, 9-prompt average
Rubric-based	2	FLAN-T5-large	question rubric, nugget rubric
Fine-tuned	2	Llama-3-8B (FT)	fine-tuned on TREC DL 2019–2022 qrels (± LLMJudge dev)
Reserved for systems	8	Llama-3-8B (FT), Llama-3 + LR, FLAN-T5 + ET	fine-tuned (3), multi-prompt + logistic regression (2), rubric / direct + ExtraTrees (3)

5.2 Evaluated Systems

The official TREC runs are not redistributed with the benchmark, so we construct 32 systems that rank the judged pool of every topic, which keeps every evaluated passage judged. Sixteen classical systems cover BM25 (two parameter

settings and an unstemmed variant), query likelihood with Dirichlet ($\mu \in \{100, 1000\}$) and Jelinek–Mercer smoothing, TF-IDF and character n-gram cosine, latent semantic analysis, RM3 pseudo-relevance feedback, a term-proximity model, three learning-to-rank models (ridge regression, LambdaMART and gradient-boosted trees) trained on the 25 disjoint development topics, and two reciprocal-rank-fusion runs. Sixteen LLM-reranking systems sort the pool by the label of one of the eight reserved pipelines and break ties by LambdaMART or BM25 scores. Under human qrels the systems span nDCG@10 from 0.263 to 0.599, and 288 of the 496 system pairs differ significantly (paired t-test over topics, $p<0.05$).

5.3 Protocol, Baselines and Metrics

All experiments use five-fold topic-level cross-fitting: the calibrator is trained on the human labels of 20 historical topics and applied to 5 new topics, and every topic is new exactly once. In calibration experiments, m random human labels per new topic are revealed ($m \in \{0, 5, 10, 20\}$) and excluded from evaluation. In review experiments, a budget B (fraction of pairs) is spent per fold; CARE uses seed sets of $\max(2, \lceil 0.02|D_q| \rceil)$ pairs and R=4 rounds. Stochastic configurations are repeated ten times and averaged. Calibration baselines are the best single judge with confusion-matrix calibration, majority voting with vote fractions as probabilities, temperature-scaled votes [24], Dawid–Skene [25], supervised Bayesian classifier combination with independent judge confusion matrices [26], stacked multinomial logistic regression over judge labels, and the same stacker with per-topic prior-shift correction [27] fitted on the seed labels. Review baselines are random review with majority-vote or best-judge labels, review of the most contested pairs (highest panel vote entropy) with majority-vote labels, which mirrors escalation rules for juries of judges [16], and random review with CARE labels. Label-level metrics are accuracy, Cohen's κ, quadratic-weighted κ (QWK), ECE with 15 bins, negative log-likelihood (NLL), Brier score and the AUROC of confidence for detecting label errors; system-level metrics are Kendall's τ, τ_{AP} and the mean absolute error (MAE) of system nDCG@10. The calibrator uses multinomial logistic regression with inverse regularization 0.05; all components run on a single CPU core in seconds per fold.

6 RESULTS

6.1 RQ1: How Inconsistent Are LLM Judges?

Table 2 isolates one factor at a time. Repeated runs of the same Llama-3-8B prompt are almost deterministic (pairwise $\kappa \geq 0.997$, flip rate $\leq 0.2\%$) and induce identical system rankings. Changing the prompt is a different story: six GPT-4o prompts agree with each other at only κ=0.631 and flip 22.2% of the graded labels, and six Llama-3-8B prompts agree at κ=0.360 with a flip rate of 45.7% and a spread of 0.73 grade points in their mean label. Running one pipeline with three LLMs flips 34.9% of the labels, and holistic versus criteria-decomposed judges flip 48.1%. The differences propagate to evaluation: system rankings induced by different prompts of one LLM correlate at τ=0.838–0.895 with each other, their agreement with the human ranking ranges from 0.548 to 0.669 for Llama-3-8B prompts, and the best system changes with the prompt. Across all 25 panel judges the mean pairwise κ is only 0.344.

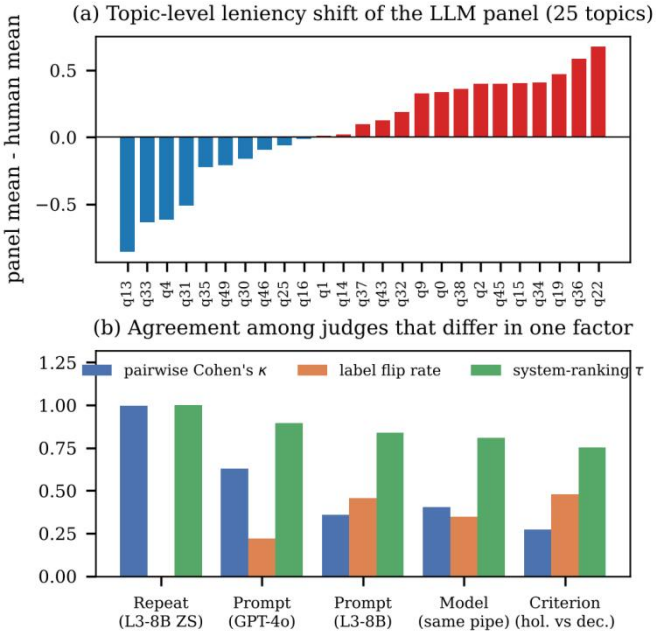


Figure 2 (a) Topic-Level Leniency of the 25-Judge Panel Relative to TREC Assessors (Panel Mean Label Minus Human Mean Label); (b) Agreement among Judges That Differ in a Single Factor

Table 2 Agreement among Judges That Differ in One Factor

Dimension	#J	κ_{pair}	Flip	Δmean	κ_H	τ_{pair}	τ_H
Repeat (L3-8B, ZS)	3	0.997	0.002	0.00	0.187–0.188	1.000	0.609
Repeat (L3-8B, CoT)	3	0.999	0.001	0.00	0.184	1.000	0.613
Prompt (GPT-4o)	6	0.631	0.222	0.26	0.239–0.286	0.895	0.625–0.669
Prompt (L3-8B)	6	0.360	0.457	0.73	0.183–0.209	0.838	0.548–0.669
Model (1 pipeline)	3	0.404	0.349	0.47	0.201–0.265	0.809	0.548–0.657
Criterion (hol. vs. dec.)	11×6	0.271	0.481	0.05	–	0.753	–

κ_{pair}, τ_{pair}: mean pairwise agreement between judges (labels, induced system rankings); Flip: fraction of pairs with different labels; Δmean: range of mean labels; κ_H, τ_H: range of agreement with TREC assessors.

Figure 2(a) shows that disagreement with humans is strongly topic-dependent: the panel's mean label exceeds the human mean by up to 0.68 grades on some topics and falls short by up to 0.86 on others. A single global correction therefore cannot remove LLM–human disagreement, which motivates the topic-adaptive calibrator.

6.2 RQ2: Can Panel Judgments Be Calibrated?

Table 3 reports calibration on new topics with ten seed labels per topic. Majority voting is the strongest unsupervised aggregator in agreement (κ=0.273), but its vote fractions are severely overconfident (ECE 0.150): pairs with near-unanimous panels are correct only about three times out of four (Figure 3). Dawid–Skene and supervised Bayesian combination are even more overconfident (ECE 0.595 and 0.419) because the judges are far from conditionally independent: judges sharing an LLM or a prompt template make correlated errors. Temperature scaling repairs calibration (ECE 0.032) but cannot change the decision boundary, so κ stays at 0.273. Stacking over raw judge labels transfers poorly to new topics (κ=0.172), and per-topic prior-shift correction recovers part of the loss (κ=0.296). CARE attains the best accuracy (0.613), κ (0.412), QWK (0.530), NLL (0.949), Brier score (0.512) and error-detection AUROC (0.728), a 51% relative κ gain over majority voting, with an ECE of 0.036 that is close to the best value.

Table 3 Calibration on New Topics (10 Seed Labels per Topic)

Method	Acc	κ	QWK	ECE↓	NLL↓	Brier↓	AUROC
Best single judge	0.499	0.233	0.475	0.132	1.149	0.629	0.601
Majority vote	0.521	0.273	0.479	0.150	1.842	0.638	0.669
Temp.-scaled votes	0.521	0.273	0.479	**0.032**	1.128	0.608	0.655
Dawid–Skene	0.392	0.174	0.457	0.595	7.911	1.199	0.696
Supervised BCC	0.481	0.247	0.485	0.419	4.177	0.909	0.655
Stacking-LR	0.477	0.172	0.404	0.084	1.179	0.644	0.657
Stacking + prior shift	0.549	0.296	0.452	0.045	1.038	0.562	0.710
CARE (ours)	**0.613**	**0.412**	**0.530**	0.036	**0.949**	**0.512**	**0.728**

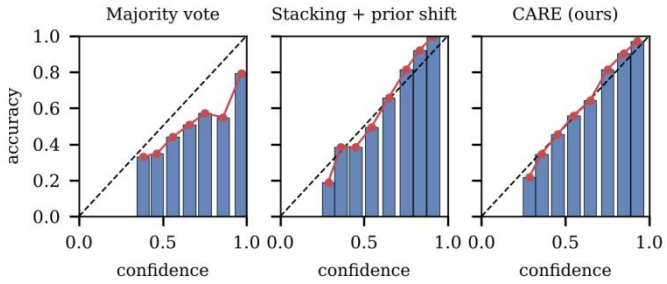


Figure 3 Reliability Diagrams on New Topics with Ten Seed Labels per Topic
Bars: accuracy per confidence bin; dashed: perfect calibration.

Table 4 dissects the calibrator. Topic adaptation is the decisive component: without it κ falls to 0.245 and does not improve with more seed labels. Replacing the 30-dimensional MDC representation by the 25 raw judge labels lowers κ from 0.412 to 0.352, and concatenating both (0.399) is also inferior, so compact consistency statistics transfer across topics better than judge identities. Among the MDC groups the topic-context features matter most (κ 0.387 without them), while the per-dimension groups are largely redundant with the panel-level statistics of a 25-judge panel ($|\Delta\kappa| \leq$ 0.003 when removed individually). A scalar ordinal offset, i.e., a pure leniency shift, is clearly worse than grade-specific biases (κ 0.326 vs. 0.412), confirming that LLMs and assessors differ in how they use individual grades. Predictive integration matters in the cold-start regime: without seed labels it reduces ECE from 0.057 to 0.030 at equal κ, whereas with ten seeds its effect is negligible. Increasing the seed labels per topic from 0 to 5, 10 and 20 raises the κ of CARE from 0.247 to 0.374, 0.412 and 0.436.

Table 4 Ablation of the CARE Calibrator

Variant	ECE (m=0)	NLL (m=0)	κ (m=10)	ECE (m=10)	NLL (m=10)
CARE (full)	0.030	1.098	0.412	0.036	0.949
MDC → raw judge labels	0.068	1.160	0.352	0.041	0.997

Variant	ECE (m=0)	NLL (m=0)	κ (m=10)	ECE (m=10)	NLL (m=10)
MDC + raw judge labels	0.045	1.122	0.399	0.030	0.966
w/o topic adaptation	0.057	1.103	0.245	0.059	1.105
w/o predictive integration	0.057	1.103	0.411	0.031	0.947
scalar ordinal offset	0.089	1.102	0.326	0.061	0.988
w/o topic-context features	–	–	0.387	0.046	0.976
panel statistics only	–	–	0.391	0.052	0.977
w/o prompt-dimension feats.	–	–	0.410	0.036	0.950
w/o model-dimension feats.	–	–	0.410	0.035	0.950
w/o criterion-dimension feats.	–	–	0.414	0.037	0.945

6.3 RQ3: Where Should Human Effort Go?

Figure 4 and Table 5 compare review policies. Random review with majority-vote labels improves label agreement linearly with the budget but barely moves the system ranking (τ from 0.617 to 0.665 at 20%), because randomly chosen labels rarely affect system differences. Reviewing the most contested pairs helps more (τ=0.758 at 20%). CARE labels alone already raise κ to 0.491 at a 10% budget, because seed and reviewed labels re-estimate each topic's grade biases. Uncertainty-driven selection gives the best label-level agreement (κ=0.511 at 10% and 0.607 at 20%), and impact-aware selection gives by far the best system-level agreement: τ=0.895 at 10% and 0.955 at 20%, versus 0.642 and 0.665 for random review with majority-vote labels. Reaching $\tau \geq 0.85$ requires 7.7% of the pairs with impact-aware CARE, 14.3% with uncertainty sampling, 28.4% with disagreement-based review and more than 50% with random review. At a 10% budget, impact-aware CARE identifies the human-best system in all ten repetitions, whereas random review with majority-vote labels does so in 6.

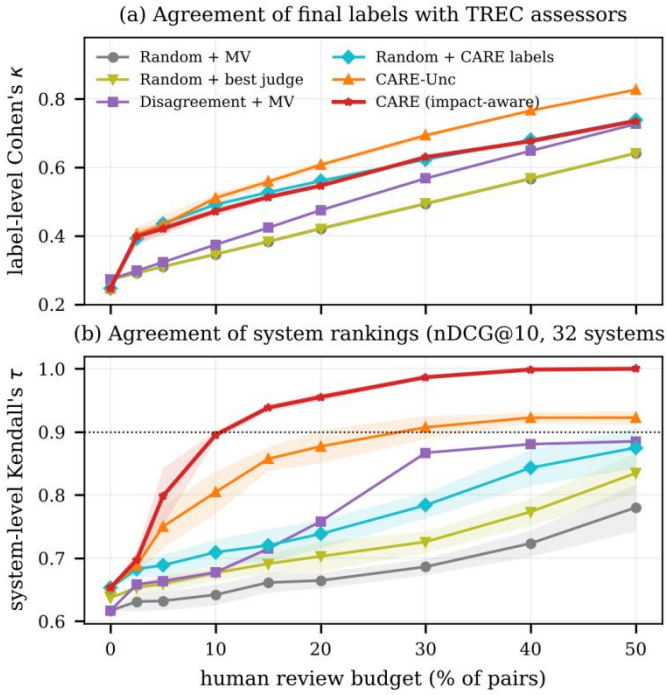


Figure 4 Label-Level κ (a) and System-Level τ (b) of the Final Qrels versus the Human Review Budget
Mean ± std over ten repetitions; dotted line: τ=0.9.

Table 5 Selective Human Review at 10% and 20% Budgets

Policy	κ@10	τ@10	τ_{AP}@10	MAE@10↓	κ@20	τ@20	B($\tau \geq .85$)
Random + MV	0.346	0.642	0.576	0.109	0.421	0.665	>50%
Random + best judge	0.346	0.677	0.626	**0.071**	0.421	0.703	>50%
Disagreement + MV	0.374	0.677	0.668	0.085	0.475	0.758	28.4%
Random + CARE labels	0.491	0.709	0.654	0.178	0.560	0.738	42.2%
CARE uncertainty	**0.511**	0.804	0.754	0.149	**0.607**	0.877	14.3%
CARE expected error	0.486	0.773	0.732	0.165	0.574	0.853	19.7%
CARE impact-aware	0.472	**0.895**	**0.813**	0.087	0.546	**0.955**	7.7%

Table 6 isolates the components of the review pipeline. Removing soft qrels lowers τ from 0.895 to 0.846 at 10% (0.955 to 0.935 at 20%) at identical κ: the expected gain is a better surrogate for system evaluation than the most likely grade. Replacing calibrated priorities and labels by impact-weighted panel disagreement with majority-vote labels costs 0.048 in τ and 0.106 in κ at 10%. Exploration seeds and multi-round updating chiefly serve label-level agreement: at 20% the

full pipeline has the highest κ (0.546 vs. 0.531 without seeds), whereas at 10% spending the entire budget on high-impact pairs yields a higher τ (0.927) at a lower κ (0.457); the number of rounds has a minor effect.

Table 6 Ablation of the Review Pipeline

Variant	κ@10	τ@10	τ_{AP}@10	κ@20	τ@20	τ_{AP}@20
CARE (full, λ=1)	0.472	0.895	0.813	0.546	0.955	0.926
w/o soft qrels	0.472	0.846	0.757	0.546	0.935	0.892
single round (R=1)	0.466	0.912	0.834	0.541	0.958	0.924
w/o exploration seeds	0.457	0.927	0.839	0.531	0.944	0.855
w/o calibration (MV labels)	0.366	0.847	0.777	0.460	0.927	0.868

Figure 5 shows the trade-off controlled by λ in (8). Pure expected-error sampling (λ=0) maximizes κ but reaches only τ=0.853 at 20%. A small impact weight (λ=0.1) loses just 0.010 in κ while achieving the highest system-level agreement of all settings (τ=0.963 at 20% and 0.898 at 10%), which makes $\lambda \approx 0.1$ a balanced default when both levels matter.

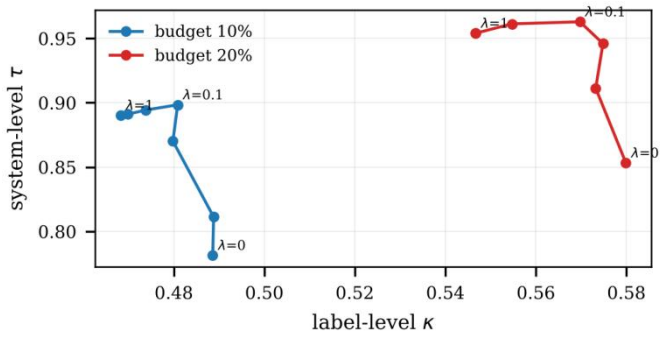


Figure 5 Trade-Off between Label-Level and System-Level Agreement Controlled by λ in (8) at 10% and 20% Budgets

6.4 RQ4: Can Calibrated Confidence Decide What Not to Review?

Figure 6 applies the acceptance rule (10) to the predictions of Table 3. With majority-vote confidence the realized error of auto-accepted labels is far above the target: 0.315 for α=0.10 and 0.461 for α=0.30. Temperature-scaled votes follow the target closely but accept few labels (30.3% at α=0.30), because vote fractions rank pairs poorly. Stacking with prior-shift correction is conservative (realized error 0.264 at 47.4% coverage). CARE keeps the realized error below the target at every α while auto-accepting 64.1% of the labels at α=0.30 (realized error 0.274) and 35.5% at α=0.20 (realized error 0.165). Calibrated confidence thus translates directly into human effort that can be saved without exceeding a prescribed error level.

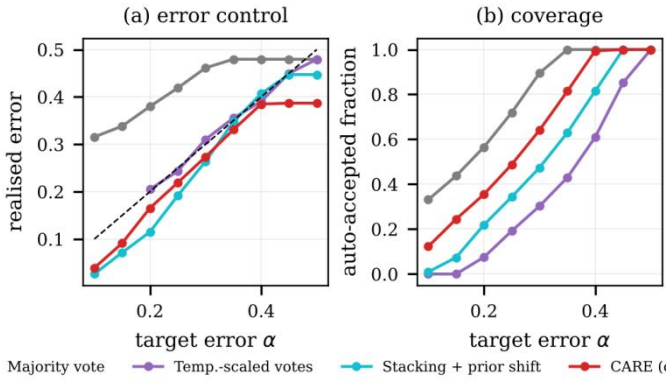


Figure 6 Risk-Targeted Auto-Acceptance: Realized Error of Accepted Labels (a) and Fraction of Labels Accepted (b) versus the Target Error α
Dashed: realized = target.

6.5 RQ5: Does CARE Depend on the Judge Panel?

Table 7 applies CARE to five panels. Within the GPT-4o panel, a practitioner who commits to one of the six prompts obtains κ between 0.239 and 0.286 and τ between 0.625 and 0.669, and majority voting over the prompts stays inside this range (τ=0.661). CARE calibration on the same six prompts reaches κ=0.401, and with a 10% review budget τ=0.883 compared with 0.694 for random review. The pattern holds for six Llama-3-8B prompts (0.590→0.907), one pipeline with three LLMs (0.642→0.898), open-weight judges (18) (0.623→0.901) and the full panel (0.648→0.890). A single LLM queried with a handful of prompts is therefore sufficient: CARE turns prompt-induced fluctuation into an uncertainty signal instead of a source of evaluation noise.

Table 7 CARE with Different Judge Panels (10% Review Budget for the Last Two Columns)

Panel	single κ	single τ	MV τ	CARE κ	Rand. τ	CARE τ
GPT-4o, 6 prompts	0.239–0.286	0.625–0.669	0.661	0.401	0.694	**0.883**
Llama-3-8B, 6 prompts	0.183–0.209	0.548–0.669	0.556	0.331	0.590	**0.907**
1 pipeline, 3 LLMs	0.201–0.265	0.548–0.657	0.621	0.378	0.642	**0.898**
Open-weight (18)	0.060–0.282	0.258–0.677	0.609	0.400	0.623	**0.901**
Full panel (25)	0.060–0.286	0.258–0.677	0.617	0.407	0.648	**0.890**

single κ/τ: range over the individual judges of the panel; CARE κ: calibrated labels with ten seeds per topic; Rand. τ: random review with majority-vote labels.

7 DISCUSSION

Implications for search and recommendation evaluation. CARE only requires graded labels from a panel and a small number of human labels, so it applies directly to recommendation settings in which LLMs grade item relevance or explanation quality [6] and to product search [5]. The results suggest three practical rules. First, prompt diversity is an inexpensive and informative source of uncertainty: one LLM queried with six prompts achieves most of the benefit of a 25-judge panel. Second, human labels are most valuable when they serve two purposes at once, namely re-estimating the grade biases of the current topic and correcting the labels that decide system comparisons. Third, confidence used for routing or acceptance must be calibrated on the target topics, because vote fractions are not probabilities.

Cost. With 4,423 pairs, a 10% budget corresponds to 442 human judgments instead of 4,423. Under that budget CARE reaches τ=0.895, a level that random review does not reach even when half of the collection is judged (τ=0.780 at 50%). Combined with risk-targeted acceptance, the framework gives evaluation teams an explicit dial between cost, label-level fidelity and system-level fidelity.

Extensions. The framework is agnostic to how judges are prompted, so listwise and pairwise LLM judges, multi-aspect recommendation criteria and streaming collections in which topics arrive over time can be integrated by extending the MDC features and the topic posterior. Combining soft qrels with prediction-powered confidence intervals [18] is a natural way to attach statistical guarantees to system-level conclusions.

8 CONCLUSION

This paper introduced CARE, an uncertainty-aware framework for LLM-based relevance evaluation that treats inconsistency across repeated runs, prompts, models and criteria as information. Using 33 real LLM label sets on TREC DL 2023 topics, we showed that prompt choice dominates the variability of LLM judgments, that LLM–human disagreement is topic-specific, and that a hierarchical topic-adaptive calibrator over multi-dimensional consistency features produces calibrated grade distributions that raise κ with TREC assessors from 0.273 to 0.412. Spending human effort according to expected error weighted by evaluation impact recovers the human system ranking with τ=0.895 at a 10% budget and τ=0.955 at 20%, and calibrated confidence enables auto-acceptance with realized error below the target. CARE thus provides a principled path toward scalable and reliable evaluation of search and recommendation systems.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Thomas P, Spielman S, Craswell N, et al. Large language models can accurately predict searcher preferences. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), 2024. DOI: 10.1145/3626772.3657707.

[2] Faggioli G, Dietz L, Clarke C L A, et al. Perspectives on large language models for relevance judgment. In: Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval (ICTIR), 2023: 39-50.

[3] MacAvaney S, Soldaini L. One-shot labeling for automatic relevance estimation. In: Proceedings of the 46th International ACM SIGIR Conference, 2023: 2230-2235.

[4] Upadhyay S, Pradeep R, Thakur N, et al. UMBRELA: UMbrela is the (open-source reproduction of the) Bing RELevance assessor. arXiv:2406.06519, 2024.

[5] Mehrdad N, Mohapatra H, Bagdouri M, et al. Large language models for relevance judgment in product search. In: Proceedings of the LLM4Eval Workshop at SIGIR, 2024. arXiv:2406.00247.

[6] Zhang X, Li Y, Wang J, et al. Large language models as evaluators for recommendation explanations. arXiv:2406.03248, 2024.

[7] Rahmani H A, Yilmaz E, Craswell N, et al. LLMJudge: LLMs for relevance judgments. In: Proceedings of the 1st Workshop on Large Language Models for Evaluation in IR (LLM4Eval) at SIGIR, 2024.

[8] Rahmani H A, Siro C, Aliannejadi M, et al. LLM4Eval: Large language model for evaluation in IR. In: Proceedings of the 47th International ACM SIGIR Conference, 2024: 3040-3043.

[9] Craswell N, Mitra B, Yilmaz E, et al. Overview of the TREC 2023 deep learning track. In: Proceedings of the 32nd Text REtrieval Conference (TREC 2023), NIST Special Publication 1328, 2023.
[10] Sclar M, Choi Y, Tsvetkov Y, et al. Quantifying language models' sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting. In: Proceedings of the International Conference on Learning Representations (ICLR), 2024.
[11] Wang P, Li L, Chen L, et al. Large language models are not fair evaluators. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL), 2024: 9440-9450.
[12] Alaofi M, Thomas P, Scholer F, et al. LLMs can be fooled into labelling a document as relevant. In: Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in IR in the Asia Pacific Region (SIGIR-AP), 2024: 32-41.
[13] Clarke C L A, Dietz L. LLM-based relevance assessment still can't replace human relevance assessment. arXiv:2412.17156, 2024.
[14] Soboroff I. Don't use LLMs to make relevance judgments. arXiv:2409.15133, 2024.
[15] Rahmani H A, Yilmaz E, Craswell N, et al. JudgeBlender: Ensembling judgments for automatic relevance assessment. arXiv:2412.13268, 2024.
[16] Verga P, Hofstätter S, Althammer S, et al. Replacing judges with juries: Evaluating LLM generations with a panel of diverse models. arXiv:2404.18796, 2024.
[17] Jung J, Brahman F, Choi Y. Trust or escalate: LLM judges with provable guarantees for human agreement. arXiv:2407.18370, 2024.
[18] Angelopoulos A N, Bates S, Fannjiang C, et al. Prediction-powered inference. Science, 2023, 382(6671): 669-674.
[19] Saad-Falcon J, Khattab O, Potts C, et al. ARES: An automated evaluation framework for retrieval-augmented generation systems. In: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), 2024.
[20] Farzi N, Dietz L. Pencils down! Automatic rubric-based evaluation of retrieve/generate systems. In: Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval (ICTIR), 2024: 175-184.
[21] Meng C, Arabzadeh N, Askari A, et al. Query performance prediction using relevance judgments generated by large language models. arXiv:2404.01012, 2024.
[22] Zheng L, Chiang W L, Sheng Y, et al. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In: Advances in Neural Information Processing Systems 36 (NeurIPS), Datasets and Benchmarks Track, 2023.
[23] Xiong M, Hu Z, Lu X, et al. Can LLMs express their uncertainty? An empirical evaluation of confidence elicitation in LLMs. In: Proceedings of the International Conference on Learning Representations (ICLR), 2024.
[24] Guo C, Pleiss G, Sun Y, et al. On calibration of modern neural networks. In: Proceedings of the 34th International Conference on Machine Learning (ICML), 2017: 1321-1330.
[25] Dawid A P, Skene A M. Maximum likelihood estimation of observer error-rates using the EM algorithm. Journal of the Royal Statistical Society: Series C (Applied Statistics), 1979, 28(1): 20-28.
[26] Kim H C, Ghahramani Z. Bayesian classifier combination. In: Proceedings of the 15th International Conference on Artificial Intelligence and Statistics (AISTATS), 2012: 619-627.
[27] Saerens M, Latinne P, Decaestecker C. Adjusting the outputs of a classifier to new a priori probabilities: A simple procedure. Neural Computation, 2002, 14(1): 21-41.
[28] Carterette B, Allan J, Sitaraman R. Minimal test collections for retrieval evaluation. In: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2006: 268-275.
[29] Yilmaz E, Aslam J A, Robertson S. A new rank correlation coefficient for information retrieval. In: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2008: 587-594.
[30] Geifman Y, El-Yaniv R. Selective classification for deep neural networks. In: Advances in Neural Information Processing Systems 30 (NeurIPS), 2017.