

Bias-Aware LLM-as-a-Judge with Position-Invariant Pairwise Evaluation and Confidence Calibration

Siyu Chen^{1*}, Matthew Hayes², Amber Long³

¹Department of Computer Science, University of Maryland, College Park, MD, USA

²Department of Computer Science and Engineering, The Ohio State University, Columbus, OH, USA

³School of Computing and Augmented Intelligence, Arizona State University, Tempe, AZ, USA

siyu.chen.works@outlook.com

Abstract

Large language models are increasingly used to compare candidate responses, but a pairwise judge can change its verdict when the same responses exchange A/B positions, and its confidence need not reflect empirical correctness. This study evaluates a position-invariant, calibration-aware pairwise framework on all 620 objectively labeled response pairs in JudgeBench. The corpus contains 528 unique questions from 17 sources and spans knowledge, reasoning, mathematics, and coding; candidate responses were generated by GPT-4o and Claude-3.5-Sonnet. Five question-grouped folds prevent repeated questions from crossing training and evaluation partitions. We compare a conventional ordered judge, swap-augmented training, an explicitly antisymmetric difference model, and two-pass position-invariant fusion. The ordered judge attains 57.26% accuracy but preserves the physical winner after an A/B swap on only 53.71% of pairs. Swap augmentation raises accuracy to 58.39% and consistency to 99.35%, while antisymmetric scoring and two-pass fusion achieve exact 100% swap consistency. The position-invariant fusion retains 58.39% accuracy, a 1.13-point gain over the ordered judge; the question-group bootstrap interval is [-2.45, 4.81] points and the exact McNemar p value is 0.597, so the accuracy change is not statistically decisive. Cross-fitted Platt scaling provides the best Brier score, 0.2367, compared with 0.2397 for raw position-invariant probabilities, and isotonic regression provides the lowest ten-bin expected calibration error, 0.0347. Confidence-based refusal improves retained-set accuracy to 69.35% at 20% coverage. The results show that order invariance can be guaranteed without reducing overall accuracy, whereas calibration and abstention must be evaluated as separate properties rather than inferred from classification accuracy.

Keywords-LLM-As-A-Judge, Pairwise Evaluation, Position Bias, Order Invariance, Confidence Calibration, Selective Prediction, Abstention, Judgebench.

A. Introduction

Large language models have moved from few-shot text generation to broadly capable assistants whose outputs support analysis, coding, planning, and domain-specific decisions. The scaling behavior documented for GPT-3 established that one model could transfer across many language tasks [1]; instruction tuning with human feedback subsequently made those capabilities easier to elicit and align with user intent [2]. GPT-4 further widened the set of tasks on which fluent, apparently reasoned answers could be produced [3]. As output quality and variety increased, evaluation became a systems problem in its own right: a useful evaluator must compare long answers, detect small but decisive errors, remain stable under irrelevant presentation changes, and expose uncertainty when the comparison is close.

LLM-as-a-judge methods address the cost of human review by asking a strong model to score or rank candidate responses. MT-Bench paired model-based judging with

multi-turn questions and community comparison [4]; G-Eval used explicit evaluation steps and form-filling prompts for natural-language generation quality [5]; and GPTScore treated task instructions as flexible evaluation criteria [6]. Human-preference platforms such as Chatbot Arena demonstrate the value of pairwise comparison at scale, while also showing that reliable model ranking depends on well-specified statistical aggregation rather than isolated votes [7]. These systems make automated evaluation practical, but they do not make every individual judgment trustworthy.

Order sensitivity is a particularly direct threat to pairwise validity. A logically valid preference should identify the same physical response as better whether that response is displayed first or second. Controlled experiments have shown that simply reversing candidates can change an LLM evaluator's decision, motivating balanced-position calibration [8]. Other audits have documented prompt sensitivity, anchoring, and inconsistent scoring distributions even when semantic quality is unchanged [9]. Position effects are consequential because

they contaminate both the predicted winner and every downstream quantity derived from it, including model win rates, reward-model targets, benchmark ranks, and confidence thresholds.

Objective labels are needed to distinguish harmless disagreement from an actual evaluation error. JudgeBench converts difficult knowledge, reasoning, mathematics, and coding problems into response pairs whose preference labels follow correctness rather than crowd taste [10]. The benchmark therefore supports direct tests of A/B reversal, physical-winner consistency, and probability calibration. Its difficulty is important: when candidate responses are both plausible and their final conclusions are close, a judge can no longer rely on surface fluency as a reliable proxy for correctness.

A second requirement is probabilistic reliability. Modern classifiers can rank examples well while remaining overconfident or underconfident, so accuracy and calibration have to be measured separately [11]. The Brier score provides a proper squared-error criterion for binary probabilities [12]. Calibration is also not confined to language evaluation: distilling human uncertainty into vision models explicitly treats disagreement as information rather than label noise [13]. For a pairwise judge, the analogous goal is not merely to output A or B, but to attach a probability whose magnitude tracks the observed chance that the selected response is correct.

Calibration becomes operational through selective prediction. A system that always emits a verdict forces low-evidence cases into the same pipeline as clear comparisons. Retrieval-augmented anomaly diagnosis has combined evidence narratives with selective refusal so that unsupported interpretations are not presented as certain [14]. Drift monitoring work similarly links uncertainty signals to safe retraining or rollback rather than treating a high model score as sufficient evidence for action [15]. These examples motivate an evaluation policy in which a calibrated judge can decline a low-confidence comparison, route it to a stronger evaluator, or request human review.

Position invariance can be viewed as a constrained fusion problem. Multisensor autonomous-driving systems combine perception streams whose identities must remain stable under changes in acquisition order [16]; uncertainty-aware late fusion makes the confidence of each branch part of the fusion rule rather than an afterthought [17]. The pairwise analogue is to evaluate both A/B and B/A presentations, map the second result back to the original physical responses, and combine logits in a way that is antisymmetric by construction. The fusion should remove presentation order while preserving evidence about which response is better.

Evidence organization and ranking provide a complementary perspective. Evidence-grounded retrieval systems for financial disclosure questions separate document acquisition from a supported answer decision [18]. Spectral rank-aggregation theory studies how noisy pairwise comparisons

can be combined into a more reliable ordering [19]. In the present setting, however, the first objective is more local: one pair should yield the same physical winner under both presentations. Once that invariance is enforced, calibration determines how strongly the fused score should be trusted.

This study asks four questions. First, how much order sensitivity appears in a conventional pairwise judge on the complete 620-pair JudgeBench corpus? Second, can swap augmentation, an antisymmetric representation, or two-pass fusion remove that sensitivity without sacrificing accuracy? Third, which cross-fitted calibration mapping provides the most useful probability estimates? Fourth, how do the conclusions vary across response generators, task families, source datasets, and candidate-similarity levels? The design treats invariance, discrimination, calibration, and selective coverage as distinct endpoints so that improvement in one cannot conceal deterioration in another.

The contribution is an end-to-end empirical audit rather than a new frontier language model. Every trainable transformation is fitted inside question-grouped folds, and every reported pair receives one out-of-fold prediction. The experiments compare six decision rules, four post-hoc calibration mappings, nine coverage points, generator and task subgroups, source-level behavior, clustered bootstrap intervals, and exact paired tests. The resulting analysis identifies a robust systems conclusion: exact order invariance is attainable at essentially unchanged accuracy, while confidence calibration and abstention remain necessary because an invariant judge can still be wrong.

B. Method

2.1 Task, corpus, and objective labels

Each observation contains a question q , two candidate responses A and B, a response-generator identifier, a source dataset, and an objective preference label y in $\{A>B, B>A\}$. The evaluation uses the two JudgeBench files in full: 350 pairs generated by GPT-4o and 270 pairs generated by Claude-3.5-Sonnet. Together they contain 620 response pairs, 528 unique question texts, and 17 source datasets. Knowledge examples are drawn from fourteen MMLU-Pro domains; reasoning and mathematics include LiveBench tasks; coding uses LiveCodeBench. Table I summarizes generator-by-family counts, and Table II reports the complete source distribution.

The positive class is $A>B$ in the original ordering. Labels are nearly balanced: 336 pairs favor A and 284 favor B. A question hash is used only as a grouping identifier. It is not a predictive feature. This distinction matters because some question texts appear with responses from both generators; keeping a repeated question in one fold prevents the evaluator from learning its label through a near duplicate. The benchmark's objective labels support error measurement directly, unlike preference-only settings in which annotator taste and factual correctness may be inseparable.

Table I. Judgebench Composition By Response Generator And Task Family

Response generator	Knowledge	Reasoning	Mathematics	Coding	Total
claude-3-5-sonnet-20240620	143	51	45	31	270
gpt-4o-2024-05-13	143	98	67	42	350
Total	286	149	112	73	620

Table II. Source-Level Dataset Statistics

Source	Family	Pairs	Unique q.	A better
livebench-math	Mathematics	90	79	61.1%
livebench-reasoning	Reasoning	149	116	54.4%
livecodebench	Coding	73	69	47.9%
mmlu-pro-biology	Knowledge	22	16	36.4%
mmlu-pro-business	Knowledge	22	22	45.5%
mmlu-pro-chemistry	Knowledge	22	20	31.8%
mmlu-pro-computer science	Knowledge	22	19	54.5%
mmlu-pro-economics	Knowledge	22	17	50.0%
mmlu-pro-engineering	Knowledge	22	18	63.6%
mmlu-pro-health	Knowledge	22	18	59.1%
mmlu-pro-history	Knowledge	22	19	54.5%
mmlu-pro-law	Knowledge	22	20	81.8%
mmlu-pro-math	Mathematics	22	19	68.2%
mmlu-pro-other	Knowledge	22	17	36.4%
mmlu-pro-philosophy	Knowledge	22	18	59.1%
mmlu-pro-physics	Knowledge	22	21	45.5%
mmlu-pro-psychology	Knowledge	22	20	63.6%

2.2 Question-grouped cross-validation

The protocol uses five-fold StratifiedGroupKFold with seed 20260317. Stratification preserves the $A > B / B > A$ balance as closely as possible, while the group constraint keeps every copy of a question in one fold. For each outer fold, the text transformation, inverse-document-frequency weights, numeric feature scaling, and classifier coefficients are fitted on the other four folds and then applied once to the held-out pairs. The implementation uses scikit-learn's sparse estimators and group-aware splitting utilities [20]. Pooled metrics concatenate the five disjoint held-out predictions, so each of the 620 labeled pairs contributes exactly once.

Calibration uses a second cross-fitting layer. For an example in outer fold k , the probability mapping is fitted only on raw out-of-fold predictions from the other four folds. The held-out fold therefore does not influence either the base judge or the calibration function that transforms its score. This arrangement is stricter than calibrating once on the same pooled predictions used for reporting and avoids optimistic reliability estimates.

2.3 Ordered representation and baselines

The ordered judge encodes the concatenation of the question and the two labeled candidates with hashed word unigrams and bigrams. Fold-local TF-IDF weighting follows the classical term-weighting principle that rare terms contribute more discriminative evidence than corpus-wide common words [21]. The text vector is augmented with structural features: response token counts, sentence counts, code-fence indicators, arithmetic-expression counts, explicit uncertainty markers, answer-option matches, lexical overlap with the question, A/B length difference, and response similarity. The classifier is an L2-regularized logistic model. Its output p_o is the estimated probability that the response displayed as A is objectively better.

Two nontrainable controls expose trivial positional and stylistic behavior. Always choose first returns A for every original and swapped presentation. Longer response selects the response with more Unicode word-and-symbol tokens. Length is included because verbose answers can appear more complete even when they contain the same or more errors. Recommendation and classification systems often use multiple lexical and behavioral signals rather than a single popularity proxy [22]; the ordered judge follows that multifeature principle while retaining a deliberately transparent linear decision surface.

2.4 Swap augmentation and antisymmetric scoring

Swap augmentation duplicates every training pair. The original record (q, A, B, y) is paired with $(q, B, A, 1-y)$. The ordered feature extractor is unchanged, but the training distribution now contains equal evidence that either display position may hold the physical winner. At evaluation time, the model is run on both presentations. If $p_o(q, B, A)$ is the probability that the swapped first response is better, its mapped probability for physical A is $1 - p_o(q, B, A)$. Swap consistency is the fraction of pairs for which the original and mapped swapped hard decisions select the same physical response.

The antisymmetric judge removes absolute position features from the learned comparison. It constructs a response-specific representation $r(q, x)$ for each candidate and uses the difference $\phi(q, A, B) = r(q, A) - r(q, B)$, followed by a zero-intercept logistic model. Swapping candidates changes ϕ to $-\phi$, so the logit changes sign exactly and $p(A > B | q, A, B) = 1 - p(B > A | q, B, A)$. This constraint guarantees physical-winner consistency independently of training balance. It is conceptually related to rank aggregation, but it produces a local pairwise probability rather than a global item ranking.

2.5 Position-invariant two-pass fusion

The proposed fusion starts from the swap-augmented ordered judge. Let z_{AB} be its logit for the original presentation and z_{BA} the logit for the swapped presentation, still expressed as the probability that the displayed first response wins. The fused logit for physical A is $z_{PI} = (z_{AB} - z_{BA})/2$, and $p_{PI} = \text{sigma}(z_{PI})$. When the inputs are swapped, the two

terms exchange places and the fused logit becomes $-z_PI$. The method therefore has exact antisymmetry even if the underlying ordered branch retains a small residual order effect.

Two-pass fusion differs from averaging raw probabilities. Logit subtraction treats equal and opposite evidence symmetrically and preserves an interpretable zero boundary. It also exposes disagreement: if both presentations favor their first position, z_AB and z_BA partially cancel rather than producing a spuriously decisive winner. Table III lists the six evaluated decision rules and their order-handling properties, while Fig. 1 summarizes the full pipeline.

Table III. Pairwise Decision Rules And Order Handling

Method	Representation	Training	Order handling
Always choose first	Position-only control	None	Always first
Longer response	Character length	None	Physical response preserved
Ordered judge	Hashed word 1-2 grams + structural features	Original order	Single pass
Swap-augmented judge	Same ordered features	Original + swapped/inverted pairs	Single pass
Antisymmetric judge	TF-IDF[QA(A)] - TF-IDF[QA(B)] + feature differences	Original pairs; zero intercept	Score sign reverses exactly
Position-invariant fusion	Original and swapped ordered logits	Swap-augmented base judge	Antisymmetric two-pass fusion

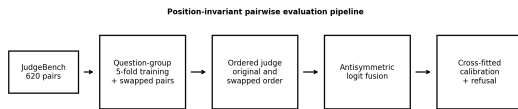


Fig. 1. *Position-Invariant Pairwise Evaluation, Calibration, And Selective-Refusal Pipeline.*

2.6 Cross-fitted confidence calibration

Four mappings transform p_PI into a probability of physical A being better. Temperature scaling divides each fold's logit by one positive scalar chosen by negative log-likelihood. Platt scaling fits a logistic mapping from the raw logit to the binary label [23]. Isotonic regression fits a monotone piecewise-constant mapping, following the broader strategy of converting arbitrary scores to empirical probabilities [24]. Beta calibration fits a logistic model over $\log(p)$ and $\log(1-p)$, allowing asymmetric corrections near the probability boundaries [25].

Calibration is assessed with Brier score, log loss, ten-bin expected calibration error (ECE), and maximum calibration

error (MCE). Accuracy and macro-F1 are also reported because a nonmonotone mapping such as isotonic regression can alter the 0.5 decision boundary. In high-stakes medical prediction, calibrated confidence determines whether an automated estimate should be reviewed rather than merely whether its ranking is correct [26]. Financial forecasting likewise combines heterogeneous uncertainty signals before committing to a directional decision [27]. These applications motivate reporting probability quality as a first-class result.

2.7 Selective refusal

For a calibrated physical-A probability p , confidence is $c=\max(p,1-p)$. At a target coverage α , the policy retains the α fraction of examples with largest c and abstains on the rest. Selective accuracy is correctness among retained pairs; selective risk is one minus that quantity. The protocol follows the risk-coverage formulation of selective classification [28] and the use of learned rejection rules in deep models [29]. It does not assign a special semantic label to abstention. Instead, abstention means that the pair requires another evaluator, external evidence, or human review.

A refusal threshold should reflect consequence and cost. Risk-aware auto-bidding explicitly optimizes a decision under both a budget and a downside constraint rather than maximizing an unconstrained point estimate [30]. The same principle applies here: low-risk deployments may accept broad coverage, whereas evaluations that determine model release, safety policy, or expensive annotation priorities can operate at lower coverage and higher retained-set accuracy.

2.8 Metrics, statistical inference, and implementation

Discrimination metrics are accuracy, macro-F1, Cohen's kappa, and ROC AUC. Order metrics are physical-winner swap consistency, mean absolute difference between original and mapped swapped probabilities, first-position preference in each presentation, and the absolute preference gap. Calibration metrics are Brier score, log loss, ECE, and MCE. Subgroup analyses use response generator, task family, source, and response-similarity tercile. Candidate similarity is cosine similarity between response TF-IDF vectors and is calculated without labels.

Paired accuracy differences are evaluated with the exact McNemar test on discordant decisions [31]. Ninety-five-percent intervals are obtained from 3,000 nonparametric bootstrap samples of complete question groups, preserving dependence among repeated-generator versions of a question in accordance with clustered resampling principles [32]. The experiment runs locally in Python with fixed seeds. The compact linear architecture also separates the statistical question from cloud-serving choices; hybrid LLM deployment studies emphasize that cost, locality, and architecture must be specified independently of model quality [33]. All tables, figures, fold predictions, calibration parameters, hashes, and logs are produced by one executable analysis program.

C. Results and Discussion

3.1 Overall accuracy and order invariance

Table IV reports pooled out-of-fold performance. Always choosing the first response reaches 54.19% because A is objectively better in 336 of 620 pairs, but its macro-F1 is only 0.351 and its probability metrics are degenerate. Selecting the longer response performs worse, at 45.00% accuracy, showing that verbosity is not a reliable correctness signal on this corpus. The ordered learned judge reaches 57.26% accuracy, 0.571 macro-F1, and 0.603 ROC AUC. Swap augmentation improves accuracy to 58.39%, while the antisymmetric judge reaches 57.58%. Position-invariant fusion matches the swap-augmented accuracy of 58.39% and produces the highest mathematics accuracy among the learned methods.

Table IV. Overall Out-Of-Fold Performance

Method	Accuracy	Macro-F1	ROC AUC	Brier	ECE
Always choose first	54.19%	0.351	0.500	0.4581	0.4581
Longer response	45.00%	0.450	0.453	0.5500	0.5500
Ordered judge	57.26%	0.571	0.603	0.2486	0.0729
Swap-augmented judge	58.39%	0.582	0.617	0.2397	0.0718
Antisymmetric judge	57.58%	0.574	0.607	0.2422	0.0669
Position-invariant fusion	58.39%	0.582	0.617	0.2397	0.0718

Accuracy alone understates the difference between the models. As Table V and Fig. 2 show, the ordered judge selects the same physical winner after a swap on only 53.71% of pairs. Its original and mapped swapped probabilities differ by 0.157 on average. Swap augmentation reduces that discrepancy to 0.0006 and raises consistency to 99.35%. The four remaining failures are cases in which a minute residual probability difference crosses 0.5. The antisymmetric judge and position-invariant fusion both achieve exactly 100% physical-winner consistency and zero mapped probability discrepancy by construction.

Table V. Position Bias And Physical-Winner Consistency

Method	Swap consistency	Mean mapped diff.	First pref. orig.	First pref. swap
Always choose first	0.00%	1.0000	100.00%	100.00%
Longer response	100.00%	0.0000	46.61%	53.39%
Ordered judge	53.71%	0.1570	51.77%	50.00%
Swap-augmented judge	99.35%	0.0006	52.58%	47.10%
Antisymmetric judge	100.00%	0.0000	52.42%	47.58%
Position-invariant fusion	100.00%	0.0000	52.58%	47.42%

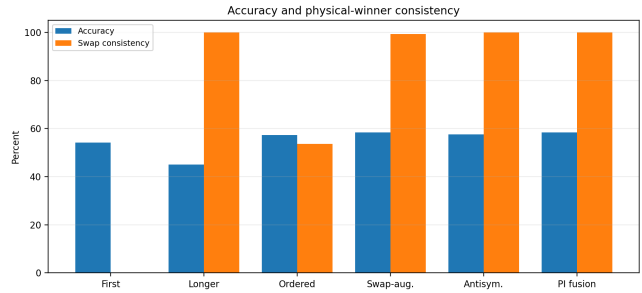


Fig. 2. Accuracy And Physical-Winner Consistency Across Decision Rules.

Figure 3 makes the mechanism visible. Ordered-judge points form a broad cloud around, rather than on, the invariance diagonal. Position-invariant probabilities lie on the diagonal up to numerical precision. This is a stronger property than a small aggregate first-position gap. For example, a method could choose the first position equally often in both presentations yet reverse the physical winner on every pair. Physical mapping directly tests the causal nuisance variable of interest: which answer was displayed first.

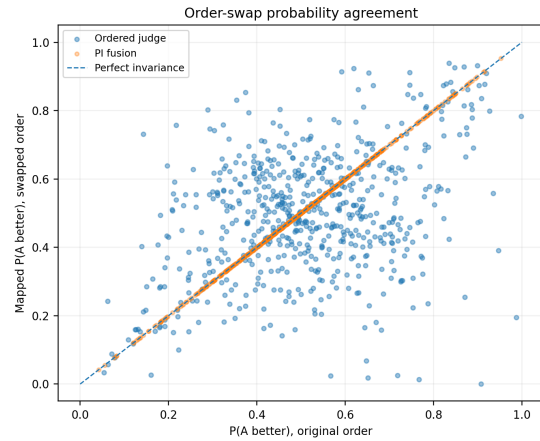


Fig. 3. Original-Order And Mapped Swapped-Order Probabilities. Position-Invariant Fusion Lies On The Invariance Diagonal.

The main benefit is therefore structural reliability, not a statistically established accuracy gain. Swap augmentation and fusion each improve accuracy by 1.13 percentage points over the ordered judge, but the question-group bootstrap interval spans -2.45 to 4.81 points for fusion. The exact McNemar test counts 61 pairs correct only under the ordered judge and 68 correct only under fusion, giving $p=0.597$. Uncertainty analyses for spectral estimators similarly emphasize that a point improvement and the uncertainty around that improvement are separate objects [34]. Here, the invariance guarantee is exact, while the accuracy difference remains uncertain.

3.2 Calibration and selective risk

Raw position-invariant probabilities have Brier score 0.2397, log loss 0.6719, and ECE 0.0718. Table VI and Fig. 4 show that every cross-fitted mapping reduces ECE. Temperature scaling lowers it to 0.0415. Platt scaling gives the best Brier

score, 0.2367, and the best log loss, 0.6653. Isotonic regression gives the lowest ECE, 0.0347, but its maximum calibration error is 0.3104 because a sparsely populated high-confidence bin is poorly estimated. Beta calibration provides the smallest MCE, 0.1211, and a Brier score within 0.00014 of Platt. No mapping dominates every criterion.

Table Vi. Cross-Fitted Calibration Results

Method	Accuracy	Brier	Log loss	ECE	MCE
Raw PI	58.39%	0.2397	0.6719	0.0718	0.1871
Temperature scaling	58.39%	0.2382	0.6682	0.0415	0.1713
Platt scaling	58.87%	0.2367	0.6653	0.0382	0.1663
Isotonic regression	58.06%	0.2377	0.6808	0.0347	0.3104
Beta calibration	58.71%	0.2369	0.6659	0.0394	0.1211

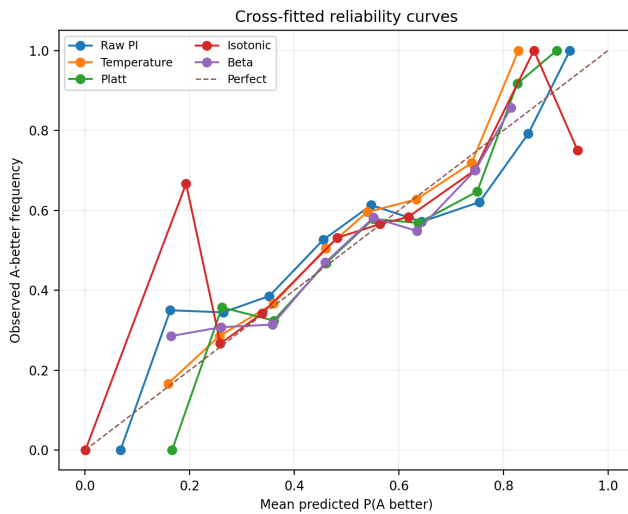


Fig. 4. Reliability curves for raw and cross-fitted calibrated probabilities.

The calibrated accuracies also require careful interpretation. Platt scaling reaches 58.87%, a three-point increase over the raw 58.39%, but its macro-F1 declines because the learned intercept changes the class balance. Isotonic regression changes several boundary decisions and reaches 58.06%. These mappings are intended to improve probability meaning, not to optimize the hard label. Credit-risk classification illustrates why this distinction matters: a decision tree can separate classes while still requiring a probability policy tailored to review and exposure [35]. Financial-service analytics similarly combine customer and risk evidence because the same score can imply different actions under different costs [36].

Table VII and Fig. 5 evaluate the action layer. With Platt scaling, retained-set accuracy rises from 58.87% at full coverage to 62.10% at 80% coverage, 63.13% at 70%, and 69.35% at 20%. Beta calibration is slightly stronger at 60%-70% coverage, reaching 64.25% at 60% and 63.36% at 70%, while Platt is strongest at 80%. The curves are not perfectly monotone because confidence ranking is imperfect and the

sample is modest. Nevertheless, the high-confidence tail is consistently more accurate than the full set.

Table VII. Selective Accuracy And Risk At Fixed Coverage

Method	Coverage	Answered	Sel. Accuracy	Sel. Risk	Mean conf.
Raw PI	100%	620	58.39%	41.61%	0.613
Raw PI	80%	496	60.08%	39.92%	0.638
Raw PI	70%	434	60.37%	39.63%	0.653
Raw PI	60%	372	61.56%	38.44%	0.669
Raw PI	40%	248	64.52%	35.48%	0.713
Raw PI	20%	124	69.35%	30.65%	0.786
Platt	100%	620	58.87%	41.13%	0.590
Platt	80%	496	62.10%	37.90%	0.609
Platt	70%	434	63.13%	36.87%	0.620
Platt	60%	372	63.98%	36.02%	0.632
Platt	40%	248	63.71%	36.29%	0.663
Platt	20%	124	69.35%	30.65%	0.719
Beta	100%	620	58.71%	41.29%	0.591
Beta	80%	496	62.10%	37.90%	0.611
Beta	70%	434	63.36%	36.64%	0.621
Beta	60%	372	64.25%	35.75%	0.633
Beta	40%	248	64.11%	35.89%	0.663
Beta	20%	124	68.55%	31.45%	0.716

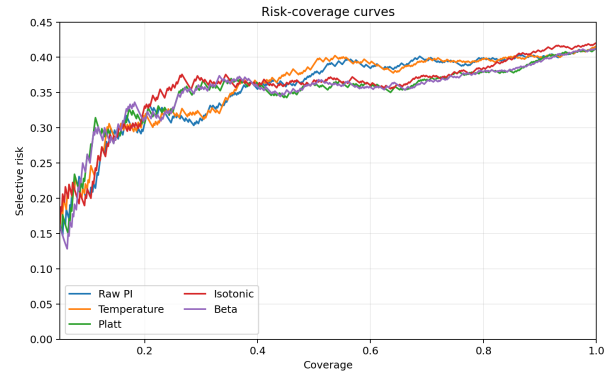


Fig. 5. Risk-coverage curves for confidence-based selective refusal.

Confidence communication is part of deployment, not merely a metric. Risk-dashboard research emphasizes visual hierarchy and explanation when uncertainty must be interpreted by institutional users [37]. A pairwise-evaluation interface should therefore display the selected response, calibrated confidence, whether both orderings agreed before fusion, and the evidence responsible for the margin. A single percentage without these provenance cues can invite false precision, particularly because the best observed full-coverage accuracy is still below 60%.

3.3 Generator, task-family, and source heterogeneity

Generator subgroups differ sharply. Table VIII shows that the ordered judge reaches 60.29% on GPT-4o-generated

response pairs and 53.33% on Claude-3.5-Sonnet pairs. Position-invariant fusion raises GPT-4o-pair accuracy to 62.86% but lowers the Claude subgroup to 52.59%. The result does not imply that one generator is intrinsically easier in general; the benchmark pairs, error types, and response styles differ. It does show that one global evaluator can hide conditional failure modes.

Table VIII. Performance By Response Generator

Generator	Method	N	Accuracy	Macro-F1	Brier
claude-3-5-sonnet	Ordered	270	53.33%	0.531	0.2560
claude-3-5-sonnet	Swap-augmented	270	52.59%	0.526	0.2453
claude-3-5-sonnet	Antisymmetric	270	52.59%	0.525	0.2485
claude-3-5-sonnet	PI fusion	270	52.59%	0.526	0.2453
gpt-4o	Ordered	350	60.29%	0.593	0.2429
gpt-4o	Swap-augmented	350	62.86%	0.625	0.2353
gpt-4o	Antisymmetric	350	61.43%	0.611	0.2373
gpt-4o	PI fusion	350	62.86%	0.625	0.2353

Task families in Table IX and Fig. 6 reinforce that conclusion. Fusion is strongest on mathematics, where it reaches 67.86%, compared with 63.39% for the ordered judge. It also improves reasoning from 56.38% to 60.40%. Knowledge remains at 55.24%, and coding declines from 57.53% to 52.05%. The 73 coding pairs are too few for a definitive architecture comparison, but the direction suggests that lexical contrast and compact structural features do not capture execution-level correctness. Retrieval-summary integration for code intelligence makes the same separation explicit: finding relevant code and explaining it are different stages [38]. A stronger coding judge should execute tests or inspect formal traces rather than infer correctness from prose alone.

TABLE IX. Accuracy By Task Family

Task family	N	Ordered	Swap aug.	Antisym.	PI fusion
Knowledge	286	55.24%	55.24%	56.64%	55.24%
Reasoning	149	56.38%	60.40%	55.03%	60.40%
Mathematics	112	63.39%	66.96%	66.07%	67.86%
Coding	73	57.53%	53.42%	53.42%	52.05%

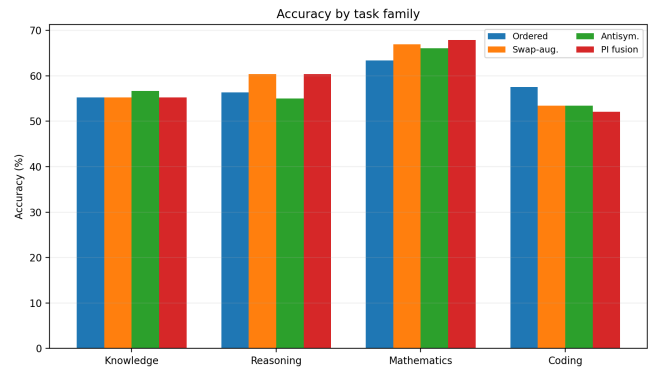


Fig. 6. Accuracy by task family and order-handling method.

Language and domain variation offer additional reasons to avoid one undifferentiated confidence threshold. Bilingual computing-education research shows that the language context of technical material can affect engagement and understanding [39]. DenseNet-based medical-image classification demonstrates a domain in which representation quality is tied to highly specific visual evidence rather than generic text similarity [40]. Autonomous-driving detection optimized with receptive-field attention likewise depends on task-specific spatial cues [41]. These studies do not provide results for JudgeBench, but they illustrate why evaluator evidence should match the structure of the task being judged.

Source-level values in Table X and Fig. 7 should be interpreted descriptively because most MMLU-Pro sources contain only 22 pairs. LiveBench mathematics is the clearest positive slice: fusion reaches 73.33%, compared with 65.56% for the ordered judge. LiveBench reasoning improves from 56.38% to 60.40%. LiveCodeBench moves in the opposite direction, from 57.53% to 52.05%. Across the small knowledge sources, fusion ranges from 27.27% on health to 68.18% on computer science. The spread indicates that calibration pooled over all sources cannot be assumed to transfer unchanged. Theory on approximately shared features predicts exactly this tension: useful structure may transfer across domains without being strictly invariant everywhere [42].

TABLE X. Source-Level Ordered And Position-Invariant Results

Source	N	Ordered acc.	PI acc.	Beta Brier
livebench-math	90	65.56%	73.33%	0.191
livebench-reasoning	149	56.38%	60.40%	0.245
livecodebench	73	57.53%	52.05%	0.258
mmlu-pro-biology	22	54.55%	54.55%	0.231
mmlu-pro-business	22	54.55%	63.64%	0.227
mmlu-pro-chemistry	22	59.09%	59.09%	0.260
mmlu-pro-computer science	22	72.73%	68.18%	0.201
mmlu-pro-economics	22	63.64%	54.55%	0.241
mmlu-pro-engineering	22	54.55%	45.45%	0.237
mmlu-pro-health	22	45.45%	27.27%	0.280

Source	N	Ordered acc.	PI acc.	Beta Brier
mmlu-pro-history	22	54.55%	59.09%	0.241
mmlu-pro-law	22	54.55%	63.64%	0.209
mmlu-pro-math	22	54.55%	45.45%	0.240
mmlu-pro-other	22	50.00%	59.09%	0.254
mmlu-pro-philosophy	22	54.55%	59.09%	0.240
mmlu-pro-physics	22	45.45%	54.55%	0.276
mmlu-pro-psychology	22	54.55%	50.00%	0.244

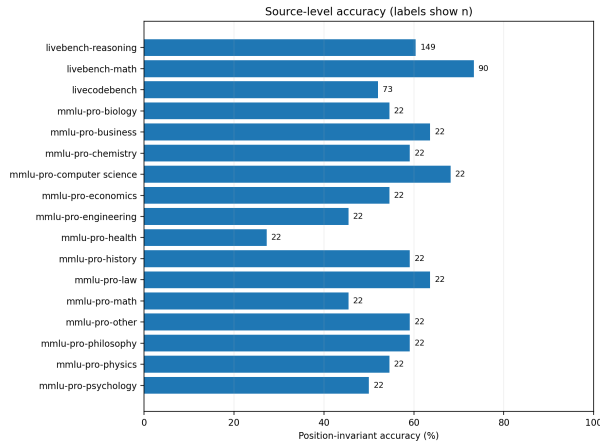


Fig. 7. Position-invariant accuracy by source; labels show the number of pairs.

3.4 Paired inference and response similarity

Table XI reports the paired comparisons. Swap augmentation, antisymmetric scoring, and fusion all have confidence intervals that include zero and McNemar p values above 0.59. These results rule out a strong claim that order control alone materially improves correctness on this sample. They also reject the opposite operational conclusion that invariance is costly: fusion attains the same best observed accuracy as swap augmentation and exceeds the ordered judge descriptively while guaranteeing exact consistency.

Table Xi. Paired Accuracy Comparisons Against The Ordered Judge

Comparison	Delta (pp)	95% cluster CI	Ordered only	Method only	McNemar p
Swap-augmented - Ordered	1.13	[-2.55, 4.82]	61	68	0.597
Antisymmetric - Ordered	0.32	[-3.17, 3.72]	58	60	0.927
PI fusion - Ordered	1.13	[-2.45, 4.81]	61	68	0.597

Response-similarity terciles in Table XII reveal where the gain occurs. On low-similarity pairs, the ordered judge leads fusion, 61.84% versus 57.97%. On medium-similarity pairs, fusion leads 58.74% versus 50.97%. On high-similarity pairs, the methods are almost tied, 58.45% versus 58.94%. Beta-

calibrated mean confidence falls from 0.606 in the low-similarity tercile to 0.572 in the high-similarity tercile, which is directionally sensible: closely matched candidates should be harder to distinguish. Yet high-similarity accuracy does not fall monotonically, confirming that lexical resemblance is only one component of difficulty.

Table Xii. Performance By Response-Similarity Tercile

Similarity	N	Cosine range	Ordered acc.	PI acc.	Beta conf.
Low	207	0.177-0.601	61.84%	57.97%	0.606
Medium	206	0.602-0.713	50.97%	58.74%	0.595
High	207	0.713-1.000	58.94%	58.45%	0.572

The medium-similarity gain is consistent with the cancellation role of two-pass fusion. When answers share much of their content but differ in one decisive claim, an ordered model can attach weight to layout, length, or answer-label position. Evaluating both orders suppresses stable position terms while retaining response-specific contrast. At very low similarity, however, the ordered model may exploit broad stylistic or content cues that the antisymmetric fusion partially shrinks. This trade-off suggests that future judges should combine exact order invariance with richer factual and executable evidence, rather than relaxing invariance to regain superficial signals.

3.5 Operational implications across evidence, monitoring, and decision systems

The results support a layered evaluation service. The base comparison should be order-invariant; calibration should be updated on recent labeled traffic; and low-confidence pairs should be escalated rather than forced. AIOps fault-analysis workflows connect detection, diagnosis, and response as separate stages [43]. IoT anomaly detection similarly distinguishes representation learning from the threshold that turns a score into an alert [44]. In an LLM evaluation service, swap consistency is a precondition check, calibrated confidence is the alert score, and the escalation target may be a verifier, a larger judge, or a human reviewer.

Operational drift requires continuous auditing. Residual-fusion models for noisy-neighbor VM degradation focus on deviations from normal behavior rather than a fixed global rule [45]. Retrieved normal prototypes have also been used to explain OpenStack failure-injection anomalies [46]. CI failure diagnosis extends the same idea from detection to evidence-backed repair recommendations [47]. Applied to LLM judges, a monitoring record should retain both order-specific logits, fused probability, calibration version, source family, and final review outcome. Drift can then be localized to a generator, task type, or probability region instead of appearing only as a falling aggregate accuracy.

Security-sensitive decisions add another reason to preserve uncertainty. Predictive DDoS mitigation uses learned risk estimates to prioritize defensive action in distributed systems [48]. A pairwise judge may similarly control consequential

choices such as which model response reaches users or which preference labels enter training. An invariant but weak judge should not be granted broad authority merely because it is stable. The measured 58%-59% full-coverage accuracy makes the need for evidence-gated escalation explicit.

Long-context evaluation should also separate retrieval from adjudication. Contract and insurance clause analysis uses long-document RAG to locate controlling text before drawing a conclusion [49]. Progressive PDF rendering prioritizes document slices so that the most useful content becomes available earlier [50], while edge-deployable video-quality predictors compress richer quality judgments into efficient signals [51]. These designs suggest a scalable judge pipeline: retrieve the decisive claim, test it with a task-specific verifier, then pass a compact evidence record to the order-invariant comparison layer. Position fusion cannot recover evidence that was never selected.

Human-facing explanations should communicate both evidence and uncertainty. Visual explanation cards for traffic-sign alerts combine localization, classification, and retrieved support instead of presenting an isolated class label [52]. The same pattern can be used for pairwise evaluation: show the disputed spans, the physical response identities, the two order-specific outcomes, the fused verdict, and the reason for any refusal. Such an interface makes position auditing visible and helps reviewers distinguish a low-margin semantic conflict from an obvious factual or executable failure.

Finally, evaluation policies are resource-allocation policies. Digital-twin dispatching combines forecast state with constrained downstream action [53]; data-driven marketing models convert predictions into enterprise decisions [54]; and constrained budgeting links macro forecasts to response models under an explicit resource limit [55]. Counterfactual learning-to-rank evaluates ranking policies from logged bandit feedback rather than treating observed clicks as unbiased labels [56]. These examples reinforce the paper's central separation: prediction, uncertainty, and action require different validation. For an LLM judge, the selected winner is the prediction, calibrated confidence is the risk estimate, and acceptance or escalation is the policy.

4. Limitations

The first limitation is model scope. The learned judges are transparent local discriminative evaluators, not frontier generative models producing free-form critiques. This design isolates the effects of representation, swapping, antisymmetry, and calibration, but it does not estimate the absolute performance of a specific commercial LLM judge. A frontier evaluator may use richer semantic evidence and may exhibit different position effects; it should nevertheless be tested with the same physical-winner protocol.

Second, JudgeBench contains 620 pairs and 528 question groups. The complete corpus supports paired analysis, but source-level estimates are noisy, especially for the fourteen

MMLU-Pro slices with 22 pairs each and for the 73 coding pairs. The clustered bootstrap accounts for repeated questions inside this corpus; it does not cover benchmark selection, future generators, multilingual prompts, or new task distributions.

Third, the benchmark contains only strict $A > B$ and $B > A$ labels. It does not test ties, partial credit, multi-dimensional rubrics, adversarial instructions embedded in candidate responses, or cases where both answers are unacceptable. A production system needs an explicit tie/invalid option and should verify that swapping preserves those outcomes as well.

Fourth, calibration is in-distribution and cross-fitted on the same benchmark family. The procedure prevents record-level leakage, but the mapping can drift when generator style, domain mix, or model competence changes. Calibration should be refitted on recent labels and reported by meaningful subgroups. The observed generator and task differences show that one global threshold is not guaranteed to preserve the same risk everywhere.

Fifth, selective accuracy remains modest. At 20% coverage, the best reported value is 69.35%, so refusal enriches for easier cases but does not create a high-assurance judge. The system is appropriate as an auditable comparison component or triage layer, not as a substitute for objective execution, domain experts, or authoritative ground truth. Position invariance prevents one class of procedural error; it does not establish semantic correctness.

D. Conclusion

This study evaluated order bias, probability calibration, and selective refusal on all 620 objectively labeled JudgeBench pairs. A conventional ordered judge reached 57.26% accuracy but retained the same physical winner after an A/B swap on only 53.71% of cases. Swap augmentation increased consistency to 99.35%; an antisymmetric difference model and two-pass logit fusion guaranteed 100% consistency. Position-invariant fusion reached 58.39% accuracy, statistically indistinguishable from the ordered model, showing that exact invariance need not impose an observed accuracy penalty.

Cross-fitted calibration improved probability quality without solving the underlying discrimination problem. Platt scaling produced the best Brier score, isotonic regression the lowest ECE, and beta calibration the smallest MCE. Selective refusal raised accuracy among retained high-confidence pairs, but even the most selective reported operating point remained below 70%. The practical conclusion is therefore layered: eliminate presentation-order dependence by construction, calibrate the resulting score on held-out data, expose disagreement and provenance, and escalate low-confidence or high-impact comparisons to evidence-based verification. Reliable LLM evaluation requires all four steps; none can be inferred from raw accuracy alone.

E. References

- [1] T. B. Brown et al., "Language models are few-shot learners," in *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877-1901, 2020.
- [2] L. Ouyang et al., "Training language models to follow instructions with human feedback," in *Advances in Neural Information Processing Systems*, vol. 35, pp. 27730-27744, 2022.
- [3] OpenAI, "GPT-4 technical report," arXiv:2303.08774, 2023.
- [4] L. Zheng et al., "Judging LLM-as-a-judge with MT-Bench and Chatbot Arena," arXiv:2306.05685, 2023.
- [5] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, and C. Zhu, "G-Eval: NLG evaluation using GPT-4 with better human alignment," in *Proc. 2023 Conf. Empirical Methods in Natural Language Processing*, pp. 2511-2522, 2023.
- [6] J. Fu, S.-K. Ng, Z. Jiang, and P. Liu, "GPTScore: Evaluate as you desire," arXiv:2302.04166, 2023.
- [7] W.-L. Chiang et al., "Chatbot Arena: An open platform for evaluating LLMs by human preference," in *Proc. 41st Int. Conf. Machine Learning*, PMLR, vol. 235, pp. 8359-8388, 2024.
- [8] P. Wang et al., "Large language models are not fair evaluators," arXiv:2305.17926, 2023.
- [9] R. Stureborg, D. Alikaniotis, and Y. Suhara, "Large language models are inconsistent and biased evaluators," arXiv:2405.01724, 2024.
- [10] S. Tan, S. Zhuang, K. Montgomery, W. Y. Tang, A. Cuadron, C. Wang, R. A. Popa, and I. Stoica, "JudgeBench: A benchmark for evaluating LLM-based judges," arXiv:2410.12784, 2024.
- [11] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. 34th Int. Conf. Machine Learning*, PMLR, vol. 70, pp. 1321-1330, 2017.
- [12] W. Brier, "Verification of forecasts expressed in terms of probability," *Monthly Weather Review*, vol. 78, no. 1, pp. 1-3, 1950. Z. S. Zhong, R. Ma, and H. Zhao, "Human-Uncertainty Distillation for Calibrated Vision Models on CIFAR-10H," *Journal of Advanced Computing Systems*, vol. 3, no. 2, pp. 77-89, Feb. 2023, doi: 10.69987/JACS.2023.30206.
- [13] J. Nie and D. Zheng, "Ambiguity-Aware HDFS Log Anomaly Detection with Retrieval-Augmented Failure Narratives and Selective Refusal," *Journal of Advanced Computing Systems*, vol. 3, no. 1, pp. 66-80, Jan. 2023, doi: 10.69987/JACS.2023.30105.
- [14] Zhang, "DriftGuard: Multi-Signal Drift Early Warning and Safe Re-Training/Rollback for CTR/CVR Models," *Journal of Advanced Computing Systems*, vol. 3, no. 7, pp. 24-40, Jul. 2023, doi: 10.69987/JACS.2023.30703.
- [15] Y. Wang, S. Du, Q. Xin, Y. He, and W. Qian, "Autonomous Driving System Driven by Artificial Intelligence Perception Fusion," *Academic Journal of Science and Technology*, vol. 9, no. 2, pp. 193-198, Feb. 2024, doi: 10.54097/e0b9ak47.
- [16] Q. Xin, "Uncertainty-Aware Late Fusion for 3D Perception (Confidence Calibration + Fusion Rule Learning)," *Journal of Technology Informatics and Engineering*, vol. 4, no. 1, pp. 215-238, 2025, doi: 10.51903/jtie.v4i1.485.
- [17] S. Zhou, Z. Li, and E. Wang, "Evidence-Grounded RAG for Tokenized Trade Receivable Disclosure QA under U.S. Capital Market Standards," *Journal of Advanced Computing Systems*, vol. 3, no. 7, pp. 41-57, Jul. 2023, doi: 10.69987/JACS.2023.30704.
- [18] Z. S. Zhong and S. Ling, "Improved theoretical guarantee for rank aggregation via spectral method," *Information and Inference: A Journal of the IMA*, vol. 13, no. 3, Art. no. iaae020, Sep. 2024, doi: 10.1093/imaia/iaae020.
- [19] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825-2830, 2011.
- [20] Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information Processing & Management*, vol. 24, no. 5, pp. 513-523, 1988. K. Xu, H. Zhou, H. Zheng, M. Zhu, and Q. Xin, "Intelligent classification and personalized recommendation of e-commerce products based on machine learning," *Applied and Computational Engineering*, vol. 64, pp. 147-153, 2024, doi: 10.54254/2755-2721/64/20241365.
- [21] C. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *Advances in Large Margin Classifiers*, A. J. Smola, P. L. Bartlett, B. Scholkopf, and D. Schuurmans, Eds. Cambridge, MA, USA: MIT Press, pp. 61-74, 1999.
- [22] B. Zadrozny and C. Elkan, "Transforming classifier scores into accurate multiclass probability estimates," in *Proc. 8th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, pp. 694-699, 2002.
- [23] M. Kull, T. M. Silva Filho, and P. Flach, "Beta calibration: A well-founded and easily implemented improvement on logistic calibration for binary classifiers," in *Proc. 20th Int. Conf. Artificial Intelligence and Statistics*, PMLR, vol. 54, pp. 623-631, 2017.
- [24] Chen, J. Xiong, Y. Wang, Q. Xin, and H. Zhou, "Implementation of an AI-based MRD Evaluation and Prediction Model for Multiple Myeloma," *Frontiers in Computing and Intelligent Systems*, vol. 6, no. 3, pp. 127-131, Jan. 2024, doi: 10.54097/zJ4MnbWW.
- [25] H. Zhou and K. Zhang, "News-Based Uncertainty and Macro-Market Fusion for VIX Direction Forecasting: Evidence from 2015-2024 FRED Panel," *Journal of Technology Informatics and Engineering*, vol. 4, no. 2, pp. 487-501, Aug. 2025, doi: 10.51903/jtie.v4i2.540.

- [26] R. El-Yaniv and Y. Wiener, "On the foundations of noise-free selective classification," *Journal of Machine Learning Research*, vol. 11, pp. 1605-1641, 2010. Y. Geifman and R. El-Yaniv, "Selective classification for deep neural networks," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [27] H. Zhang, "Risk-Aware Budget-Constrained Auto-Bidding under First-Price RTB: A Distributional Constrained Deep Reinforcement Learning Framework," *Journal of Advanced Computing Systems*, vol. 4, no. 6, pp. 30-47, Jun. 2024, doi: 10.69987/JACS.2024.40603.
- [28] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," *Psychometrika*, vol. 12, no. 2, pp. 153-157, 1947.
- [29] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. New York, NY, USA: Chapman & Hall, 1993.
- [30] Q. Xin, "Hybrid Cloud Architecture for Efficient and Cost-Effective Large Language Model Deployment," *Journal of Information Systems and Informatics*, vol. 7, no. 3, pp. 2182-2195, Sep. 2025, doi: 10.51519/journalisi.v7i3.1170.
- [31] Z. S. Zhong and S. Ling, "Uncertainty Quantification of Spectral Estimator and MLE for Orthogonal Group Synchronization," *arXiv:2408.05944*, Aug. 2024.
- [32] Q. Xin, R. Song, Z. Wang, Z. Xu, and F. Zhao, "Enhancing Bank Credit Risk Management Using the C5.0 Decision Tree Algorithm," *Journal of Computer Technology and Applied Mathematics*, vol. 1, no. 4, pp. 100-107, Nov. 2024, doi: 10.5281/zenodo.14032041.
- [33] T. Yang, Q. Xin, X. Zhan, S. Zhuang, and H. Li, "Enhancing Financial Services through Big Data and AI-Driven Customer Insights and Risk Analysis," *Journal of Knowledge Learning and Science Technology*, vol. 3, no. 3, pp. 53-62, Jul. 2024, doi: 10.60087/jklst.vol3.n3.p53-62.
- [34] Z. Li, S. Zhou, and Z. Zhou, "Financial Risk Dashboard Design for Institutional RWA Investors: Visual Hierarchy, Chart Comprehension, and Explainability in FinChart-Bench," *International Journal of Graphic Design*, vol. 3, no. 1, pp. 196-210, May 2025, doi: 10.51903/ijgd.v3i1.3715.
- [35] Y. Li, "Findable then Explainable: Retrieval-Summary Integration for Code Intelligence on a Lightweight CodeSearchNet Subset," *Journal of Advanced Computing Systems*, vol. 4, no. 7, pp. 65-82, Jul. 2024, doi: 10.69987/JACS.2024.40706.
- A. G. S. Raj, H. Zhang, V. Abhyankar, S. Mukherjee, E. Zhang, J. Williams, R. Halverson, and J. M. Patel, "Impact of bilingual CS education on student learning and engagement in a data structures course," in *Proc. 19th Koli Calling Int. Conf. Computing Education Research*, Art. no. 24, pp. 1-10, 2019, doi: 10.1145/3364510.3364518.
- [36] Z. Zhong, M. Zheng, H. Mai, J. Zhao, and X. Liu, "Cancer image classification based on DenseNet model," *Journal of Physics: Conference Series*, vol. 1651, no. 1, Art. no. 012143, Nov. 2020, doi: 10.1088/1742-6596/1651/1/012143.
- [37] Z. Ling, Q. Xin, Y. Lin, G. Su, and Z. Shui, "Optimization of autonomous driving image detection based on RFACnv and triplet attention," *Applied and Computational Engineering*, vol. 77, pp. 210-217, 2024, doi: 10.54254/2755-2721/77/2024MA0067.
- [38] Z. S. Zhong, X. Pan, and Q. Lei, "Bridging Domains with Approximately Shared Features," in *Proc. 28th Int. Conf. Artificial Intelligence and Statistics*, PMLR, vol. 258, pp. 559-567, 2025.
- [39] Y. He, Y. Pan, Y. Wang, S. Du, and Q. Xin, "Intelligent Fault Analysis with AIOps Technology," *Journal of Theory and Practice of Engineering Science*, vol. 4, no. 1, pp. 94-100, Feb. 2024, doi: 10.53469/jtpes.2024.04(01).13.
- [40] Q. Xin, Z. Xu, L. Guo, F. Zhao, and B. Wu, "IoT traffic classification and anomaly detection method based on deep autoencoders," *Applied and Computational Engineering*, vol. 69, pp. 64-70, 2024, doi: 10.54254/2755-2721/69/20241511.
- [41] Nie and D. Zheng, "Noisy-Neighbor-Aware VM Degradation Risk Modeling with Unsupervised Residual Fusion," *Journal of Advanced Computing Systems*, vol. 4, no. 4, pp. 112-123, Apr. 2024, doi: 10.69987/JACS.2024.40409.
- [42] Q. Xin, "Explaining OpenStack Failure-Injection Log Anomalies with Retrieved Normal Prototypes," *Emerging Information Science and Technology*, vol. 6, no. 2, pp. 125-146, Nov. 2025, doi: 10.18196/eist.v6i2.31232.
- [43] H. Zhang, "LLM-Driven CI Failure Diagnosis and Automated Repair: From GitHub Actions Logs to Patch Recommendation," *Journal of Technology Informatics and Engineering*, vol. 4, no. 1, pp. 190-214, Apr. 2025, doi: 10.51903/jtie.v4i1.484.
- [44] B. Wang, Y. He, Z. Shui, Q. Xin, and H. Lei, "Predictive optimization of DDoS attack mitigation in distributed systems using machine learning," *Applied and Computational Engineering*, vol. 64, pp. 94-99, May 2024, doi: 10.54254/2755-2721/64/20241350.
- [45] S. Zhou, Z. Li, and E. Wang, "Long-Document RAG for Contractual and Insurance Clause Analysis in Receivables RWA Structures," *Journal of Advanced Computing Systems*, vol. 4, no. 8, pp. 88-104, Aug. 2024, doi: 10.69987/JACS.2024.40810.
- [46] H. Wang, Y. Ren, and X. Chang, "Layout-Aware Progressive PDF Rendering: AI Prioritization of PDF Slices to Reduce Time-to-Functional-First-Frame on FUNSD," *Journal of Technology Informatics and Engineering*, vol. 4, no. 2, pp. 425-446, Aug. 2025, doi: 10.51903/jtie.v4i2.523.

- [47] B. Zhou, H. Wang, and X. Chang, "Distilling VMAF into an Edge-Deployable Quality Predictor: A Pilot Shot-Level Proxy with LLM-Ready Quality Tokens," *Journal of Technology Informatics and Engineering*, vol. 4, no. 2, pp. 447-463, Aug. 2025, doi: 10.51903/jtie.v4i2.522.
- [48] R. Ma, L. Zhang, and T. Song, "Computer-Vision-Informed Visual Explanation Cards for Autonomous-Driving Traffic-Sign Alerts: Localization, Classification, and Retrieved Evidence on GTSDb," *International Journal of Graphic Design*, vol. 3, no. 2, pp. 455-473, Oct. 2025, doi: 10.51903/ijgd.v3i2.3991.
- [49] Zhang, R. Ma, and P. Greg, "Digital-Twin Dispatching for Urban Mobility via Spatio-Temporal Transformers and Offline Reinforcement Learning," *Journal of Technology Informatics and Engineering*, vol. 4, no. 2, pp. 337-363, Aug. 2025, doi: 10.51903/jtie.v4i2.501.
- [50] J. Wang, Q. Xin, Y. Liu, J. Wang, and T. Yang, "Predicting Enterprise Marketing Decision Making with Intelligent Data-Driven Approaches," *Journal of Industrial Engineering and Applied Science*, vol. 2, no. 3, pp. 12-19, Jun. 2024, doi: 10.5281/zenodo.11357252.
- [51] Y. Lu, H. Zhou, and Y. Zhang, "A Constrained, Data-Driven Budgeting Framework Integrating Macro Demand Forecasting and Marketing Response Modeling," *Journal of Technology Informatics and Engineering*, vol. 4, no. 3, pp. 493-520, Dec. 2025, doi: 10.51903/jtie.v4i3.466.
- [52] H. Zhang, "Counterfactual Learning-to-Rank for Ads: Off-Policy Evaluation on the Open Bandit Dataset," *Journal of Advanced Computing Systems*, vol. 5, no. 12, pp. 1-11, Dec. 2025, doi: 10.69987/JACS.2025.51201.