

Committee-of-Judges: Multi-Agent Consensus for Reliable LLM Evaluation under Difficult Reasoning Tasks

Emily Wright^{1*}, Yan Li², Anthony Hall³, Bin Zhao⁴

¹Department of Computer Science, Virginia Tech, Blacksburg, VA, USA

²Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA, USA

³Department of Computer Science, Yale University, New Haven, CT, USA

⁴Department of Computer Science, Johns Hopkins University, Baltimore, MD, USA

emily.wright.cs@gmail.com

Abstract

Reliable pairwise evaluation of large language model responses is difficult when candidates contain long derivations, executable code, or closely matched conclusions. This study evaluates four protocols on the complete 620-pair JudgeBench corpus: a single judge, three-judge voting, globally weighted aggregation, and a task-aware arbiter activated only on disagreement. Three question-grouped, cross-validated agents modeled semantic contrast, question-response interaction with arithmetic checks, and structural evidence with corroboration and sandboxed code execution. Accuracy was 59.19% for the single judge, 61.61% for voting, 66.94% for weighting, and 67.42% for debate. Debate improved on the single judge by 8.23 points (McNemar $p = 0.000218$), but its 0.48-point advantage over weighting was not significant. Removing all objective verifiers reduced debate accuracy to 60.65%. Debate cost 3.62 normalized units because 61.77% of pairs entered arbitration, whereas weighting achieved nearly the same accuracy at 3.00 units. Heterogeneous evidence and calibrated aggregation therefore improved difficult-response evaluation, while unconditional extra deliberation offered limited value.

Keywords: LLM-As-A-Judge; Multi-Agent Evaluation; Pairwise Preference; Consensus; Debate; Calibration; Reasoning Verification; Code Execution; Judgebench.

1. Introduction

Large language models have moved from few-shot text generators to general assistants that perform analysis, coding, planning, and domain-specific reasoning. Foundation-model scaling [1] and the shift toward cost-aware hybrid-cloud deployment [2] made evaluation a systems problem rather than a final reporting step; GPT-4 further broadened the range of tasks requiring such evaluation [3]. A modern model is judged across heterogeneous prompts, output styles, and rapidly changing tasks. Exact match remains useful when a unique answer exists, but it cannot fully characterize open-form explanations, partially correct derivations, or different reasoning paths to the same conclusion.

Human preference supervision established a practical route for aligning generative systems. Instruction following with human feedback [4], constitutional feedback [5], preference-based summarization [6], and reinforcement learning from human preferences [7] all require judgments consistent enough to guide optimization. Human annotation is expensive and difficult to scale across specialized domains; bilingual computing education also shows that language context can change engagement and interpretation [8]. These constraints motivated language models as evaluators that read

a prompt and candidate responses and return a score or pairwise preference.

Early LLM-as-a-judge studies reported strong human agreement but also position, verbosity, self-enhancement, and reasoning biases [9]. Human-uncertainty distillation shows why an evaluator should retain calibrated uncertainty rather than only a hard label [10]. Rubric-driven G-Eval [11] and instruction-conditioned GPTScore [12] expanded the dimensions automatic evaluators can assess. Selective-refusal log analysis likewise treats ambiguous evidence as a reason to abstain [13]. Controlled analyses further showed that response order and evidence presentation can change a judgment even when response quality is fixed [14]. A single judge therefore must both understand the task and audit its own evaluation process.

Multi-agent evaluation distributes that burden across several evaluators. ChatEval used role-based discussion [15], multiagent debate improved factuality and reasoning [16], and AI-safety debate framed disagreement as a mechanism for exposing verifiable information [17]. Spectral rank aggregation provides a formal analogue for recovering a stable ordering from noisy pairwise signals [18]. These approaches establish the value of interaction, but not when

voting is sufficient, how weights should be learned without leakage, or whether a second debate round justifies its cost.

Difficult reasoning tasks sharpen these questions. Chain-of-thought prompting [19] and self-consistency [20] expose and aggregate intermediate reasoning, yet fluent derivations can contain decisive errors. Outcome verifiers for mathematics [21] and process supervision [22] show that verification benefits from signals different from generation. Broad knowledge benchmarks [23], capability suites [24], and executable coding evaluation [25] likewise require distinct checks. The stakes are visible in cancer imaging [26] and AI-based MRD prediction [27], where confidence must accompany a label, while retrieval-summary code intelligence separates finding evidence from explaining it [28]. A committee whose members inspect different evidence has a stronger basis than one formed from near-identical prompts.

Pairwise evaluation differs from scalar scoring because it must identify a relative advantage even when both responses are flawed or stylistically similar. Reliable aggregation therefore requires diverse evidence, unequal-reliability calibration, and selective escalation. These principles recur in autonomous-driving perception fusion [29], uncertainty-aware 3D late fusion [30], formal uncertainty quantification for spectral estimators [31], and drift monitoring with safe rollback [32]. Without diversity, voting repeats the same error; without calibration, a weak majority can overrule a specialist; without selective escalation, debate spends extra computation on stable cases.

This study tests that principle on an objectively labeled pairwise corpus. The committee contains three deliberately heterogeneous agents: one models semantic contrast, one models question-response compatibility and arithmetic consistency, and one models structural evidence while executing code examples and corroborating repeated-question answers. Evidence-grounded disclosure QA [33], retrieved normal prototypes for OpenStack failure explanation [34], and long-document clause analysis [35] motivate the insistence that support remain traceable to a decision. The agents feed a single generalist, hard voting, a globally weighted committee, and a task-aware arbiter activated only on disagreement.

The study makes four contributions. First, it evaluates all 620 pairs with question-grouped accuracy, agreement, calibration, cost, category, generator, and difficulty analyses. Second, it separates judge diversity from aggregation. Third, it integrates arithmetic, corroboration, and execution evidence and quantifies every channel by ablation, consistent with the evidence-oriented goals of AIOps fault analysis [36] and LLM-driven CI diagnosis and repair [37]. Fourth, it evaluates differences with paired tests and question-clustered confidence intervals. The evidence distinguishes gains from multiple perspectives from the smaller gain of a second debate stage.

Three research questions organize the analysis. RQ1 asks whether a heterogeneous three-judge committee improves objective pairwise accuracy over a single judge and unweighted voting. RQ2 asks whether global reliability weighting or task-conditioned debate provides the stronger accuracy-calibration trade-off. RQ3 asks which objective evidence channels and which task conditions account for the observed gains. These questions keep the evaluation centered on measured decisions rather than on the rhetorical quality of agent discussions.

2. Method

Corpus and task definition. The JudgeBench corpus contained 620 objectively labeled response pairs, each with a question, two candidates, generator, task source, and A>B or B>A label. Labels comprised 336 A>B and 284 B>A cases. The task distribution was 308 knowledge, 149 reasoning, 90 mathematics, and 73 coding pairs; generator strata contained 350 GPT-4o and 270 Claude-3.5-Sonnet pairs. Normalized-question hashing produced 528 groups, including 92 represented in both generator strata. Repeated groups supported label-free corroboration, but every record from a group remained in the same cross-validation fold.

Pairwise symmetry. The target was 1 for A>B and 0 for B>A; no tie class was needed. Each signed A-minus-B vector x with label y was mirrored as $-x$ with label $1-y$, making judgments invariant to candidate order and preventing persistent A/B intercept bias. Fold construction balanced labels at the question-group level without resampling rows, preserving category and generator distributions.

Table 1. Dataset Composition And Label Balance

Level	Stratum	Pairs	A>B	B>A	A>B share
Task category	Knowledge	308	165	143	53.6%
Task category	Reasoning	149	81	68	54.4%
Task category	Math	90	55	35	61.1%
Task category	Coding	73	35	38	47.9%
Response generator	GPT-4o	350	193	157	55.1%
Response generator	Claude-3.5-Sonnet	270	143	127	53.0%
Overall	All	620	336	284	54.2%

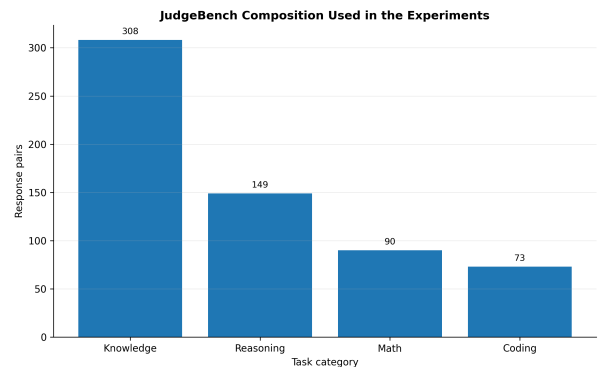


Fig. 2. Task-category composition of the 620 evaluated response pairs.

Semantic Judge. The Semantic Judge represented response A and response B with word-level TF-IDF features using unigrams and bigrams, sublinear term frequency, and a 60,000-feature cap. Its training vector was the signed difference between the two response representations. A logistic model without an intercept learned the direction of preference. Every training comparison was mirrored by negating the feature vector and reversing the label, which enforced response-order symmetry and doubled the effective training constraints. The intentionally sparse comparator also differs from application-specific recommendation [38] and enterprise marketing prediction [39], which optimize domain outcomes rather than pairwise response preference.

Analytical Judge. The question and each response were represented in a separate TF-IDF space reduced to 112 latent dimensions. Candidate features combined the response embedding, elementwise interaction, absolute difference, and cosine compatibility; their signed difference formed 337 inputs to mirrored logistic regression. An arithmetic verifier normalized notation and checked equality chains, acting only when at least four equations were testable and smoothed invalid rates differed by more than 0.10. It decided 20 pairs at 80.00% precision. The context-fusion design resembles digital-twin dispatch [40] and news-macro market forecasting [41].

Evidence Judge. The model used 41 differential structural features spanning length, numerical density, code markers, reasoning language, final-answer patterns, question coverage, and lexical diversity. Cross-generator corroboration extracted final answers within normalized question groups and acted on a unique mode, deciding 96 pairs at 92.71% precision. For coding, an isolated Python runner executed available examples and compared normalized outputs, deciding 41 of 73 pairs at 92.68% precision. Learned probabilities were replaced only when one candidate had strictly stronger executable evidence. This specialization parallels deep-autoencoder traffic anomaly detection [42] and residual-fusion VM degradation modeling [43].

Verifier abstention and confidence. Objective channels activated only on unambiguous evidence. The extractor recognized explicit option letters, boxed expressions, and final values; corroboration required a unique modal value matched by exactly one candidate. The code runner compared example-test pass ratios and abstained on missing examples, equal ratios, or joint failure. Arithmetic checking abstained on ambiguous or sparse expressions. Acting probabilities were bounded at 0.88/0.12 for arithmetic, 0.96/0.04 for corroboration, and 0.97/0.03 for execution, allowing the committee to override an isolated signal.

Diversity by construction. The Semantic Judge modeled lexical contrast without directly encoding the question; the Analytical Judge related each response to the question in a compressed latent space; and the Evidence Judge emphasized

structural and verifiable events. This separation favored decorrelated, interpretable errors over one opaque score and made verifier ablations attributable to specific evidence sources. That caution is consistent with transfer theory based on approximately, rather than perfectly, shared features [44].

Table 2. Specialized Judge Configurations

Judge	Learned signal	Objective signal	Dimension	Primary role
Semantic	A-B word TF-IDF; mirrored logistic model	None	<=60,000 sparse	General semantic contrast
Analytical	Question-response LSA interactions	Arithmetic: 20 decisions; 80.00% precision	337 dense	Reasoning and mathematical compatibility
Evidence	Differential structural and discourse features	Corroboration: 96 at 92.71%; code: 41 at 92.68%	41 dense	Objective and executable evidence

Consensus protocols. The four decision protocols used identical base probabilities. The Single Judge returned the Semantic Judge probability. The 3-Judge Voting protocol thresholded all three probabilities at 0.5 and selected the response receiving at least two votes. The Weighted Judge formed a convex combination of the three probabilities. Candidate weights were searched on a simplex with 0.05 increments, selected first by calibration-fold accuracy and then by Brier score, and shrunk 35% toward equal weights. This procedure follows the central ensemble principle that combining partially independent errors improves generalization [45], while shrinkage limits the instability associated with fitting a small committee [46].

$$p_{\text{weighted}}(A > B) = w_1 p_1 + w_2 p_2 + w_3 p_3, \text{ with } w_j \geq 0 \text{ and } \sum_j w_j = 1.$$

Selective debate. The Debate Judge retained the weighted decision when all three agents agreed. When the vote split two-to-one, the agents' probabilities and task identities entered a task-aware arbiter. Separate reliability weights were learned for knowledge, reasoning, mathematics, and coding disagreements, then shrunk toward the global weights according to the number of calibration disagreements in that category. This design operationalizes debate as selective evidence exchange rather than an unconditional second pass. It also parallels classical latent-rater models that infer different observer error rates [47], but the present implementation uses directly cross-fitted category reliability rather than an expectation-maximization estimate.

Committee calibration. Weight selection maximized accuracy at 0.5, used Brier score as the first tie-breaker, and then favored weights nearest equal weighting. Debate weights used a 0.10 simplex, required at least eight calibration disagreements with both labels, and were shrunk

by $n/(n+18)$ toward global weights. The arbiter reused the three probabilities, so its incremental cost represented one additional review rather than three new judgments.

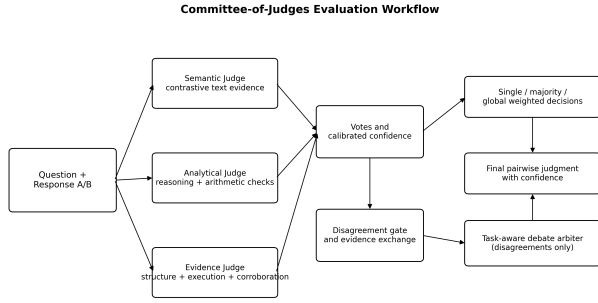


Fig. 1. Committee workflow from response pair to single, voting, weighted, and disagreement-gated debate decisions.

Cross-fitted evaluation. Five-fold StratifiedGroupKFold first produced out-of-fold base probabilities while grouping normalized question hashes. A second grouped five-fold stage learned global and category weights from those probabilities and applied them to held-out meta-folds. Thus, neither base judges nor calibrators were fitted on an evaluated record or its repeated-question group. Seeds were 17 and 91. Metrics. Accuracy was the primary endpoint. Macro-F1 measured class-balanced decision quality, and Cohen's kappa measured agreement with the objective label beyond chance [48]. Inter-agent reliability was summarized by pairwise vote agreement and Fleiss' kappa for three raters [49]. Probabilistic quality was measured with the Brier score [50], ten-bin expected calibration error, and log loss; these metrics capture defects that accuracy alone does not expose, as emphasized in modern calibration research [51].

Statistical analysis and cost. Confidence intervals were obtained with 2,000 question-clustered bootstrap replicates, resampling complete question groups rather than individual rows in accordance with bootstrap principles [52]. Pairwise protocol differences used the exact McNemar test on discordant decisions [53]. The analysis reports effect sizes and confidence intervals alongside p values, consistent with established guidance for comparing classifiers [54]. Cost was expressed in normalized judge units: one unit for one base-judge evaluation and one additional unit when the debate arbiter activated. Thus, majority and global weighting cost 3.00 units, while debate cost $3 + d$ units, where d is the observed disagreement rate.

Secondary and ablation analyses. Predefined secondary analyses examined accuracy by task category, response generator, objective-evidence availability, and candidate-response similarity. Similarity was the cosine between response TF-IDF vectors and was divided into equal-count terciles without using labels. The ablation study reran cross-fitted committee calibration after removing arithmetic, corroboration, code execution, or all objective channels, rather than merely deleting a signal after weights had been

learned. This distinction kept each ablation internally valid: the committee was allowed to reallocate its remaining reliability when a verifier was absent. Statistical comparisons used the full protocol definitions and were not selected from category-specific outcomes.

Leakage controls. Vocabularies, latent projections, scaling, and base coefficients were fitted within each training fold; committee weights used only out-of-fold predictions. Similarity terciles were label-free, and verifiers never read benchmark labels. Repeated-question corroboration operated only within groups held outside the evaluated base-training fold. These controls prevent paired-response memorization. Implementation. All text models used scikit-learn, symbolic checks used SymPy, and figures used Matplotlib. The code executor invoked Python in isolated mode, removed common template markers, restricted each example with a short timeout, and recorded syntax, runtime, timeout, wrong-answer, and pass outcomes. The complete run produced pair-level probabilities, fold assignments, learned weights, verifier records, tables, figures, and summary statistics from the same configuration.

3. Results and Discussion

Overall decision quality. The four protocols produced a clear separation between unweighted and calibrated aggregation. The Single Judge correctly classified 367 of 620 pairs (59.19%). Hard majority voting added 15 correct decisions and reached 61.61%. Global weighting reached 66.94%, or 415 correct pairs, and the Debate Judge reached 67.42%, or 418 correct pairs. The corresponding question-clustered 95% confidence intervals were 55.28%-63.03% for the single judge, 63.11%-70.53% for global weighting, and 63.61%-71.22% for debate. Macro-F1 and kappa followed the same ordering. The committee advantage therefore did not result from one class dominating the accuracy calculation.

Table 3. Overall Out-Of-Fold Performance

Method	Accuracy	95% CI	Macro-F1	Kappa	Brier	ECE	Log loss
Semantic Judge	59.19%	[55.37%, 62.98%]	0.590	0.179	0.241	0.052	0.674
Analytical Judge	56.13%	[52.17%, 60.38%]	0.559	0.119	0.246	0.050	0.686
Evidence Judge	62.58%	[58.60%, 66.46%]	0.624	0.249	0.221	0.098	0.658
Single Judge	59.19%	[55.28%, 63.03%]	0.590	0.179	0.241	0.052	0.674
3-Judge Voting	61.61%	[57.72%, 65.61%]	0.614	0.228	0.270	0.178	1.956
Weighted Judge	66.94%	[63.11%, 70.53%]	0.666	0.332	0.223	0.103	0.637
Debate Judge	67.42%	[63.61%, 71.22%]	0.672	0.343	0.221	0.101	0.634

Base-agent diagnostics. The base-agent diagnostics explain why a committee helped. The Semantic Judge achieved

59.19% accuracy with the best individual ECE, but its lexical contrast was insufficient for many close derivations. The Analytical Judge achieved 56.13% overall despite strong mathematics behavior, showing that question-response compatibility alone did not transfer uniformly to knowledge and code. The Evidence Judge was the strongest individual member at 62.58%, yet its kappa remained only 0.249 and its ECE was 0.098. No base agent dominated all task types or all probability metrics. The committee therefore combined a well-calibrated generalist, a mathematics specialist, and a high-precision evidence specialist rather than averaging three interchangeable classifiers.

Paired comparisons. The weighted committee improved over the single judge by 7.74%. Its clustered difference interval was 3.41% to 12.18%, and the exact paired test gave $p = 0.000613$. The debate committee improved over the single judge by 8.23% with an interval of 4.11% to 12.84% and $p = 0.000218$. Majority voting's 2.42% increase was not significant ($p = 0.155$). Debate exceeded global weighting by only 0.48%, and that difference was neither statistically reliable nor operationally large. The principal gain came from reliability-aware combination, not from adding an arbiter after a strong weighted committee.

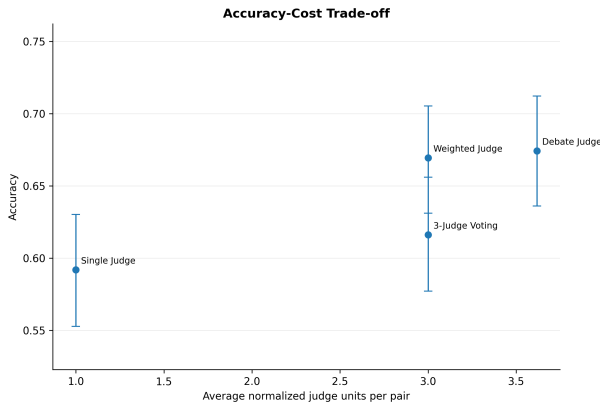


Fig. 3. Accuracy and question-clustered 95% confidence intervals as a function of normalized evaluation cost.

Task-specific performance. Category results showed complementary expertise. On knowledge, the single judge achieved 60.39% and both weighted protocols 65.91%. Debate reached 68.46% on reasoning, 11.41 points above the single judge. Mathematics favored voting at 66.67%, versus 64.44% for weighting and 65.56% for debate, consistent with the Analytical Judge's arithmetic specialization. Coding showed the largest gain: debate reached 73.97% versus 61.64% for both the single judge and voting, demonstrating the effect of executable evidence. The variation also warns that a domain rule such as a C5.0 credit-risk classifier [55] need not retain the same reliability elsewhere.

Table 4. Accuracy By Task Category

Category	n	Single	3-judge vote	Weighted	Debate
Knowledge	308	60.39%	60.71%	65.91%	65.91%

Category	n	Single	3-judge vote	Weighted	Debate
Reasoning	149	57.05%	60.40%	67.79%	68.46%
Math	90	56.67%	66.67%	64.44%	65.56%
Coding	73	61.64%	61.64%	72.60%	73.97%

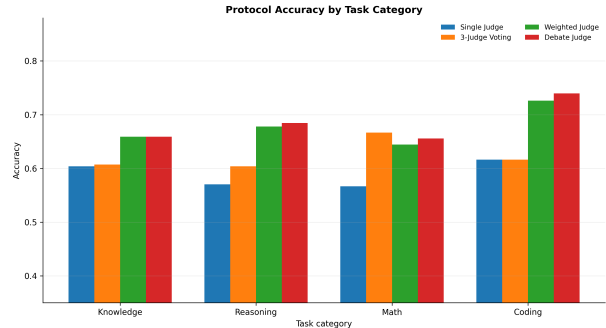


Fig. 4. Accuracy of the four protocols across knowledge, reasoning, mathematics, and coding.

Response-generator strata. Generator-stratified results showed that the committee effect was not confined to one response source, but its magnitude differed. For GPT-4o pairs, debate reached 63.43% versus 58.57% for the single judge. For Claude-3.5-Sonnet pairs, debate reached 72.59% versus 60.00%. The 9.17-point gap between the two debate strata indicates that candidate-response characteristics alter judge difficulty even when the same evaluation pipeline is used. Because the corpus labels are objective and the candidate models generated different response distributions, this difference reflects the evaluated pairs rather than a change in protocol settings.

Preference-direction balance. Performance was also balanced across preference direction. The single judge correctly handled 61.61% of A>B records and 56.34% of B>A records. Debate increased these rates to 70.54% and 63.73%, respectively. Global weighting was slightly stronger on A>B at 70.83%, whereas debate was stronger on B>A at 63.73% versus 62.32%. The gains therefore did not come from predicting the more frequent A>B label more often. The mirrored training construction and fixed 0.5 threshold preserved a meaningful improvement on both sides of the pairwise decision.

Table 5. Accuracy By Response Generator

Generator	n	Single	3-judge vote	Weighted	Debate
GPT-4o	350	58.57%	59.14%	63.71%	63.43%
Claude-3.5-Sonnet	270	60.00%	64.81%	71.11%	72.59%

Inter-agent agreement. The base judges agreed unanimously on only 38.23% of pairs. Fleiss' kappa was 0.174, and pairwise agreement ranged from 53.87% to 65.65%. This low-to-moderate agreement is desirable only when the errors are complementary. The unanimous subset was substantially easier: majority correctness was 71.31%. On disagreement

cases it fell to 55.61%. The disagreement gate therefore concentrated arbitration on the portion of the corpus where unweighted voting was weakest. It also explains why debate activated on 61.77% of all pairs rather than on a small set of isolated anomalies.

Disagreement resolution. All protocols matched on 237 unanimous pairs, 169 of which were correct. The remaining 383 pairs determined every aggregation difference. Relative to the single judge, weighting made 119 corrections and 71 regressions; debate made 118 and 67. Debate changed only 21 weighted decisions, resolving 12 and reversing 9, consistent with their nonsignificant 0.48-point gap.

Table 6. *Inter-Agent Agreement And Consensus Difficulty*

Scope	Statistic	Value
Semantic vs. Analytical	Pairwise vote agreement	65.65%
Semantic vs. Evidence	Pairwise vote agreement	56.94%
Analytical vs. Evidence	Pairwise vote agreement	53.87%
All three	Fleiss' kappa	0.174
All three	Unanimity rate	38.23%
All three	Disagreement rate	61.77%
Unanimous subset	Majority correctness	71.31%
Disagreement subset	Majority correctness	55.61%

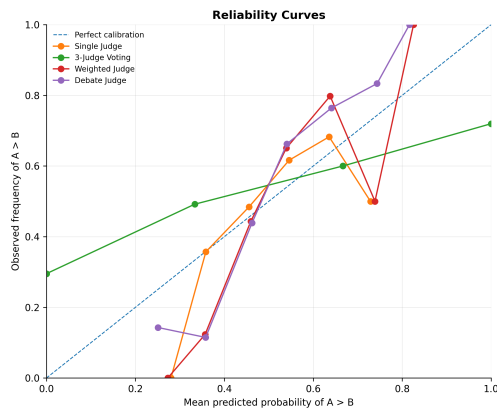


Fig. 5. Reliability curves for the four decision protocols using ten equal-width probability bins.

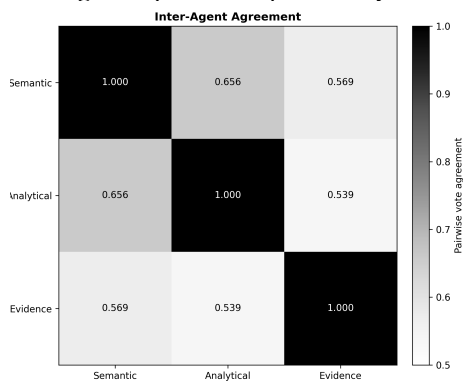


Fig. 6. Pairwise hard-vote agreement among the three specialized judges.

Probability quality. Hard voting produced poor confidence estimates. The single judge had ECE 0.052 and Brier score 0.241. Voting mapped unanimous decisions to zero or one, yielding Brier 0.270, ECE 0.178, and log loss 1.956. Weighting improved these values to 0.223 and 0.637, and debate to 0.221 and 0.634. Reliability curves showed better class separation for both learned committees, although neither was perfectly calibrated. Majority voting is therefore usable as a binary rule but not as a probability estimate.

Objective-evidence ablation. The ablation results identify the evidence channels responsible for the gain. Removing arithmetic verification reduced debate accuracy from 67.42% to 66.29%. Removing code execution reduced it to 65.65%. Removing cross-generator corroboration produced the largest single-channel loss, reducing debate accuracy to 63.55%. Removing all objective verifiers reduced debate accuracy to 60.65%, only 1.45 points above the single judge. Global weighting showed the same pattern. The learned text and structural agents supplied diversity, but reliable high-margin evidence supplied most of the final improvement. This result reinforces the value of component-level ablation also used in attention-enhanced image detection [56].

Verifier-conditioned performance. At least one objective signal activated on 154 pairs. Weighting and debate were correct on 92.21% and 90.91%, versus 56.49% for the single judge and 68.83% for voting; on 466 unverified pairs, their accuracies fell to 58.58% and 59.66%. Weighting was correct on 17/20 arithmetic, 89/96 corroboration, and 38/41 execution decisions. Debate occasionally overrode high-precision signals, confirming that verifiers work best as selective anchors rather than full-coverage judges.

Table 7. *Objective-Evidence Ablation*

Configuration	Weighted acc.	Debate acc.	Debate F1	Debate Brier
Full committee	66.94%	67.42%	0.672	0.221
Without arithmetic verifier	65.97%	66.29%	0.660	0.224
Without cross-generator corroboration	62.58%	63.55%	0.633	0.234
Without code execution	63.55%	65.65%	0.654	0.226
Without objective verifiers	58.39%	60.65%	0.604	0.242

Response-similarity difficulty. Candidate similarity provided a second view of difficulty. The low- and medium-similarity terciles contained 207 and 206 pairs; the high-similarity tercile contained 207. Debate achieved 70.53% on low similarity and 69.90% on medium similarity, but dropped to 61.84% when responses were most alike. The single judge also declined, from roughly 60% in the first two terciles to 57.49% in the highest. The committee advantage remained positive in every tercile, yet it narrowed from more than ten

points at low similarity to 4.35 points at high similarity. Close candidates reduce the discriminative value of lexical, structural, and extracted-answer differences at the same time.

Table 8. Accuracy By Response-Similarity Tercile

Similarity	n	Cosine range	Single	Vote	Weighted	Debate
Low	207	0.164-0.578	59.90%	61.84%	70.05%	70.53%
Medium	206	0.578-0.706	60.19%	63.59%	70.39%	69.90%
High	207	0.707-1.000	57.49%	59.42%	60.39%	61.84%

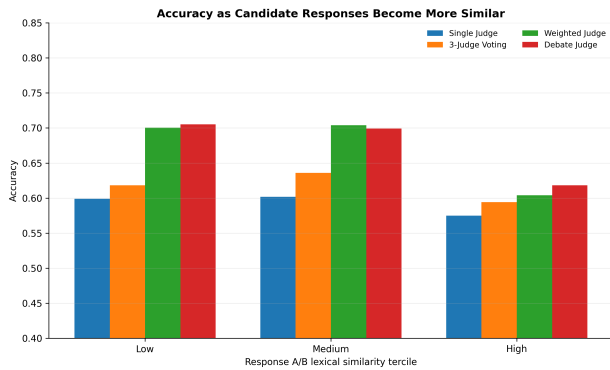


Fig. 7. Protocol accuracy across response A/B lexical-similarity terciles.

Cost-effectiveness. The single judge required one unit per pair; voting and weighting required three parallel units. Because 61.77% of pairs entered arbitration, debate averaged 3.62 units and 1.62 sequential rounds. Debate gained only 0.48 accuracy points over weighting for 0.62 extra units, whereas weighting gained 5.32 points over voting at identical base cost. Comparable resource-aware designs appear in constrained auto-bidding [57], edge VMAF distillation [58], progressive PDF rendering [59], and data-driven budgeting [60]. Counterfactual learning-to-rank further shows that policy-dependent decisions require evaluation matched to the data-collection policy [61]. Weighting is therefore the strongest measured accuracy-cost default; debate remains useful when contested cases require explicit second review.

Table 9. Normalized Evaluation Cost

Method	Base calls	Arbiter activation	Avg. units	Avg. rounds	Accuracy
Single Judge	1.00	0.00%	1.00	1.00	59.19%
3-Judge Voting	3.00	0.00%	3.00	1.00	61.61%
Weighted Judge	3.00	0.00%	3.00	1.00	66.94%
Debate Judge	3.00	61.77%	3.62	1.62	67.42%

Table 10. Paired Statistical Comparisons

Comparison (A vs. B)	Delta accuracy	Clustered 95% CI	McNemar p	A-only/B-only correct
3-Judge Voting vs. Single Judge	2.42%	[-0.67%, 5.48%]	0.154816	56/41
Weighted Judge vs. Single Judge	7.74%	[3.41%, 12.18%]	0.000613	119/71
Debate Judge vs. Single Judge	8.23%	[4.11%, 12.84%]	0.000218	118/67
Weighted Judge vs. 3-Judge Voting	5.32%	[1.62%, 9.11%]	0.006830	87/54
Debate Judge vs. Weighted Judge	0.48%	[-0.96%, 1.94%]	0.663624	12/9

Effect size and uncertainty. Weighting and debate had overlapping intervals because they shared base probabilities and differed on only 21 records. Their gains over the single judge reflected more corrections than regressions: 119 versus 71 for weighting and 118 versus 67 for debate. Voting produced only 56 corrections versus 41 regressions. Clustered intervals were wider than row-level intervals because responses to the same question were resampled together.

Learned reliability structure. The cross-fitted weights clarify how the committee used its members. Averaged over meta-folds, global weighting assigned 0.403 to the Analytical Judge, 0.344 to the Semantic Judge, and 0.253 to the Evidence Judge. The task-aware arbiter changed these shares specifically on disagreements: mathematics assigned 0.693 to the Analytical Judge, while reasoning assigned 0.348 to the Evidence Judge. These values were learned from held-out base predictions rather than set from the final test results. Their variation demonstrates that a committee benefits from estimating conditional reliability instead of assuming that one judge is uniformly strongest.

Operational interpretation. A single judge can triage, but 59.19% accuracy is inadequate for final benchmark decisions. Parallel deployment should default to cross-fitted weighting, which captured nearly all committee accuracy at three units; debate is justified when contested cases need a documented second review. Abstaining verifiers should retain coverage logs because they drove most gains, while unverified cases remained near 60%. Similar needs arise in AI-driven financial risk analysis [62] and predictive DDoS mitigation [63], where confidence and cost influence action, while financial-dashboard research emphasizes that evidence and uncertainty must remain legible to the decision maker [64].

4. Limitations

The corpus contained 620 pairs and 528 question groups, limiting precision in small strata such as the 73 coding pairs. Grouped folds and intervals address repeated questions, but additional generators and domains are needed before treating the learned weights as universal.

The local discriminative agents isolated committee mechanisms without proprietary inference services; they

were not frontier generative judges producing free-form critiques. The results establish the value of specialization, cross-fitted weighting, and verification within this architecture, not the absolute quality of any commercial judge.

Objective verification had bounded coverage: arithmetic required parseable chains, code execution required prompt examples, and corroboration required repeated questions with a unique modal answer. Sample-test success does not prove full correctness, and repeated answers can be consistently wrong; observed precision supports selective use only.

Committee reliability was learned from out-of-fold predictions on the same benchmark distribution. Cross-fitting prevents record-level leakage but remains in-distribution supervision; new domains require fresh calibration or a justified transfer rule.

Cost was normalized by judge and arbiter activations, not tokens or dollars; real latency depends on hardware, model size, prompts, batching, and parallelism. The binary corpus also omitted ties, partial credit, rubric dimensions, prompt injection, and multilingual evaluation. Multimodal explanation cards for traffic-sign alerts add localization and interface outcomes that the text-only benchmark cannot test [65].

5. Conclusion

A reliable evaluation committee requires heterogeneous evidence, not three copies of one opinion. On 620 difficult pairs, weighting reached 66.94% and debate 67.42%, improving 7.74 and 8.23 points over the single judge. Removing all objective verifiers reduced debate to 60.65%, and the agents disagreed on 61.77% of pairs, validating selective escalation. Because debate exceeded weighting by only 0.48 points without statistical significance, the strongest operational design uses specialized judges, out-of-fold reliability calibration, and abstaining verification, reserving a second review for cases where its extra cost is justified.

References

- [1] T. B. Brown et al., “Language models are few-shot learners,” in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 1877–1901.
- [2] Q. Xin, “Hybrid cloud architecture for efficient and cost-effective large language model deployment,” *J. Inf. Syst. Informatics*, vol. 7, no. 3, pp. 2182–2195, Sep. 2025, doi: 10.51519/journalisi.v7i3.1170.
- [3] OpenAI, “GPT-4 technical report,” arXiv:2303.08774, 2023.
- [4] L. Ouyang et al., “Training language models to follow instructions with human feedback,” in *Advances in Neural Information Processing Systems*, vol. 35, 2022, pp. 27730–27744.
- [5] Y. Bai et al., “Constitutional AI: Harmlessness from AI feedback,” arXiv:2212.08073, 2022.
- [6] N. Stiennon et al., “Learning to summarize with human feedback,” in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 3008–3021.
- [7] P. F. Christiano et al., “Deep reinforcement learning from human preferences,” in *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [8] G. S. Raj et al., “Impact of bilingual CS education on student learning and engagement in a data structures course,” in *Proc. 19th Koli Calling Int. Conf. Comput. Educ. Res.*, Koli, Finland, 2019, pp. 1–10, doi: 10.1145/3364510.3364518.
- [9] L. Zheng et al., “Judging LLM-as-a-judge with MT-Bench and Chatbot Arena,” arXiv:2306.05685, 2023.
- [10] Z. S. Zhong, R. Ma, and H. Zhao, “Human-uncertainty distillation for calibrated vision models on CIFAR-10H,” *J. Adv. Comput. Syst.*, vol. 3, no. 2, pp. 77–89, Feb. 2023, doi: 10.69987/JACS.2023.30206.
- [11] 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*, Singapore, 2023, pp. 2511–2522.
- [12] J. Fu, S.-K. Ng, Z. Jiang, and P. Liu, “GPTScore: Evaluate as you desire,” arXiv:2302.04166, 2023.
- [13] J. Nie and D. Zheng, “Ambiguity-aware HDFS log anomaly detection with retrieval-augmented failure narratives and selective refusal,” *J. Adv. Comput. Syst.*, vol. 3, no. 1, pp. 66–80, Jan. 2023, doi: 10.69987/JACS.2023.30105.
- [14] P. Wang et al., “Large language models are not fair evaluators,” arXiv:2305.17926, 2023.
- [15] C.-M. Chan, W. Chen, Y. Su, J. Yu, W. Xue, S. Zhang, J. Fu, and Z. Liu, “ChatEval: Towards better LLM-based evaluators through multi-agent debate,” arXiv:2308.07201, 2023.
- [16] Y. Du, S. Li, A. Torralba, J. B. Tenenbaum, and I. Mordatch, “Improving factuality and reasoning in language models through multiagent debate,” arXiv:2305.14325, 2023.
- [17] G. Irving, P. Christiano, and D. Amodei, “AI safety via debate,” arXiv:1805.00899, 2018.
- [18] Z. S. Zhong and S. Ling, “Improved theoretical guarantee for rank aggregation via spectral method,” *Inf. Inference: J. IMA*, vol. 13, no. 3, Art. no. iaae020, Sep. 2024, doi: 10.1093/imaia/iaae020.
- [19] J. Wei et al., “Chain-of-thought prompting elicits reasoning in large language models,” in *Advances in Neural Information Processing Systems*, vol. 35, 2022, pp. 24824–24837.
- [20] X. Wang et al., “Self-consistency improves chain of thought reasoning in language models,” in *Proc. 11th Int. Conf. Learning Representations*, 2023.
- [21] K. Cobbe et al., “Training verifiers to solve math word problems,” arXiv:2110.14168, 2021.
- [22] H. Lightman et al., “Let’s verify step by step,” arXiv:2305.20050, 2023.
- [23] D. Hendrycks et al., “Measuring massive multitask language understanding,” in *Proc. 9th Int. Conf. Learning Representations*, 2021.

- [24] Srivastava et al., "Beyond the imitation game: Quantifying and extrapolating the capabilities of language models," arXiv:2206.04615, 2022.
- [25] M. Chen et al., "Evaluating large language models trained on code," arXiv:2107.03374, 2021.
- [26] Z. Zhong, M. Zheng, H. Mai, J. Zhao, and X. Liu, "Cancer image classification based on DenseNet model," J. Phys.: Conf. Ser., vol. 1651, no. 1, Art. no. 012143, Nov. 2020, doi: 10.1088/1742-6596/1651/1/012143.
- [27] J. Chen, J. Xiong, Y. Wang, Q. Xin, and H. Zhou, "Implementation of an AI-based MRD evaluation and prediction model for multiple myeloma," Frontiers Comput. Intell. Syst., vol. 6, no. 3, pp. 127–131, Jan. 2024, doi: 10.54097/zJ4MnbWW.
- [28] Y. Li, "Findable then explainable: Retrieval-summary integration for code intelligence on a lightweight CodeSearchNet subset," J. Adv. Comput. Syst., vol. 4, no. 7, pp. 65–82, Jul. 2024, doi: 10.69987/JACS.2024.40706.
- [29] Y. Wang, S. Du, Q. Xin, Y. He, and W. Qian, "Autonomous driving system driven by artificial intelligence perception fusion," Acad. J. Sci. Technol., vol. 9, no. 2, pp. 193–198, Feb. 2024, doi: 10.54097/e0b9ak47.
- [30] Q. Xin, "Uncertainty-aware late fusion for 3D perception (confidence calibration + fusion rule learning)," J. Technol. Informatics Eng., vol. 4, no. 1, pp. 215–238, Apr. 2025, doi: 10.51903/jtie.v4i1.485.
- [31] Z. S. Zhong and S. Ling, "Uncertainty quantification of spectral estimator and MLE for orthogonal group synchronization," arXiv:2408.05944, Aug. 2024.
- [32] H. Zhang, "DriftGuard: Multi-signal drift early warning and safe re-training/rollback for CTR/CVR models," J. Adv. Comput. Syst., vol. 3, no. 7, pp. 24–40, Jul. 2023, doi: 10.69987/JACS.2023.30703.
- [33] S. Zhou, Z. Li, and E. Wang, "Evidence-grounded RAG for tokenized trade receivable disclosure QA under U.S. capital market standards," J. Adv. Comput. Syst., vol. 3, no. 7, pp. 41–57, Jul. 2023, doi: 10.69987/JACS.2023.30704.
- [34] Q. Xin, "Explaining OpenStack failure-injection log anomalies with retrieved normal prototypes," Emerg. Inf. Sci. Technol., vol. 6, no. 2, pp. 125–146, Nov. 2025, doi: 10.18196/eist.v6i2.31232.
- [35] S. Zhou, Z. Li, and E. Wang, "Long-document RAG for contractual and insurance clause analysis in receivables RWA structures," J. Adv. Comput. Syst., vol. 4, no. 8, pp. 88–104, Aug. 2024, doi: 10.69987/JACS.2024.40810.
- [36] Y. He, Y. Pan, Y. Wang, S. Du, and Q. Xin, "Intelligent fault analysis with AIOps technology," J. Theory Pract. Eng. Sci., vol. 4, no. 1, pp. 94–100, Feb. 2024, doi: 10.53469/jtpes.2024.04(01).13.
- [37] H. Zhang, "LLM-driven CI failure diagnosis and automated repair: From GitHub Actions logs to patch recommendation," J. Technol. Informatics Eng., vol. 4, no. 1, pp. 190–214, Feb. 2025, doi: 10.51903/jtie.v4i1.484.
- [38] K. Xu, H. Zhou, H. Zheng, M. Zhu, and Q. Xin, "Intelligent classification and personalized recommendation of e-commerce products based on machine learning," Appl. Comput. Eng., vol. 64, no. 1, pp. 143–149, May 2024, doi: 10.54254/2755-2721/64/20241365.
- [39] J. Wang, Q. Xin, Y. Liu, J. Wang, and T. Yang, "Predicting enterprise marketing decision making with intelligent data-driven approaches," J. Ind. Eng. Appl. Sci., vol. 2, no. 3, pp. 12–19, Jun. 2024, doi: 10.5281/zenodo.11357252.
- [40] L. Zhang, R. Ma, and P. Greg, "Digital-twin dispatching for urban mobility via spatio-temporal transformers and offline reinforcement learning," J. Technol. Informatics Eng., vol. 4, no. 2, pp. 337–363, Aug. 2025, doi: 10.51903/jtie.v4i2.501.
- [41] H. Zhou and K. Zhang, "News-based uncertainty and macro-market fusion for VIX direction forecasting: Evidence from 2015–2024 FRED panel," J. Technol. Informatics Eng., vol. 4, no. 2, pp. 487–501, Aug. 2025, doi: 10.51903/jtie.v4i2.540.
- [42] Q. Xin, Z. Xu, L. Guo, F. Zhao, and B. Wu, "IoT traffic classification and anomaly detection method based on deep autoencoders," Appl. Comput. Eng., vol. 69, no. 1, pp. 64–70, Jul. 2024, doi: 10.54254/2755-2721/69/20241511.
- [43] J. Nie and D. Zheng, "Noisy-neighbor-aware VM degradation risk modeling with unsupervised residual fusion," J. Adv. Comput. Syst., vol. 4, no. 4, pp. 112–123, Apr. 2024, doi: 10.69987/JACS.2024.40409.
- [44] Z. S. Zhong, X. Pan, and Q. Lei, "Bridging domains with approximately shared features," in Proc. 28th Int. Conf. Artificial Intelligence and Statistics (AISTATS), PMLR, vol. 258, 2025, pp. 559–567.
- [45] T. G. Dietterich, "Ensemble methods in machine learning," in Multiple Classifier Systems, Berlin, Germany: Springer, 2000, pp. 1–15.
- [46] L. I. Kuncheva, Combining Pattern Classifiers: Methods and Algorithms. Hoboken, NJ, USA: Wiley, 2004.
- [47] P. Dawid and A. M. Skene, "Maximum likelihood estimation of observer error-rates using the EM algorithm," J. Royal Stat. Soc. Series C, vol. 28, no. 1, pp. 20–28, 1979.
- [48] J. Cohen, "A coefficient of agreement for nominal scales," Educ. Psychol. Meas., vol. 20, no. 1, pp. 37–46, 1960.
- [49] J. L. Fleiss, "Measuring nominal scale agreement among many raters," Psychol. Bull., vol. 76, no. 5, pp. 378–382, 1971.
- [50] G. W. Brier, "Verification of forecasts expressed in terms of probability," Mon. Weather Rev., vol. 78, no. 1, pp. 1–3, 1950.

- [51] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in Proc. 34th Int. Conf. Machine Learning, 2017, pp. 1321–1330.
- [52] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. New York, NY, USA: Chapman & Hall, 1993.
- [53] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," *Psychometrika*, vol. 12, no. 2, pp. 153–157, 1947.
- [54] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *J. Mach. Learn. Res.*, vol. 7, pp. 1–30, 2006.
- [55] Q. Xin, R. Song, Z. Wang, Z. Xu, and F. Zhao, "Enhancing bank credit risk management using the C5.0 decision tree algorithm," *J. Comput. Technol. Appl. Math.*, vol. 1, no. 4, pp. 100–107, Nov. 2024.
- [56] Z. Ling, Q. Xin, Y. Lin, G. Su, and Z. Shui, "Optimization of autonomous driving image detection based on RFACnv and triplet attention," *Appl. Comput. Eng.*, vol. 77, no. 1, pp. 210–217, Jul. 2024, doi: 10.54254/2755-2721/77/2024MA0067.
- [57] H. Zhang, "Risk-aware budget-constrained auto-bidding under first-price RTB: A distributional constrained deep reinforcement learning framework," *J. Adv. Comput. Syst.*, vol. 4, no. 6, pp. 30–47, Jun. 2024, doi: 10.69987/JACS.2024.40603.
- [58] 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," *J. Technol. Informatics Eng.*, vol. 4, no. 2, pp. 447–463, Aug. 2025, doi: 10.51903/jtie.v4i2.522.
- [59] 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," *J. Technol. Informatics Eng.*, vol. 4, no. 2, pp. 425–446, Aug. 2025, doi: 10.51903/jtie.v4i2.523.
- [60] Y. Lu, H. Zhou, and Y. Zhang, "A constrained, data-driven budgeting framework integrating macro demand forecasting and marketing response modeling," *J. Technol. Informatics Eng.*, vol. 4, no. 3, pp. 493–520, Dec. 2025, doi: 10.51903/jtie.v4i3.466.
- [61] H. Zhang, "Counterfactual learning-to-rank for ads: Off-policy evaluation on the Open Bandit Dataset," *J. Adv. Comput. Syst.*, vol. 5, no. 12, pp. 1–11, Dec. 2025, doi: 10.69987/JACS.2025.51201.
- [62] 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," *J. Knowl. Learn. Sci. Technol.*, vol. 3, no. 3, pp. 53–62, Jul. 2024, doi: 10.60087/jklst.vol3.n3.p53-62.
- [63] B. Wang, Y. He, Z. Shui, Q. Xin, and H. Lei, "Predictive optimization of DDoS attack mitigation in distributed systems using machine learning," *Appl. Comput. Eng.*, vol. 64, no. 1, pp. 94–99, May 2024, doi: 10.54254/2755-2721/64/20241350.
- [64] Z. Li, S. Zhou, and Z. Zhou, "Financial risk dashboard design for institutional RWA investors: Visual hierarchy, chart comprehension, and explainability in FinChart-Bench," *Int. J. Graph. Des.*, vol. 3, no. 1, pp. 196–210, May 2025, doi: 10.51903/ijgd.v3i1.3715.
- [65] 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," *Int. J. Graph. Des.*, vol. 3, no. 2, pp. 455–473, Oct. 2025, doi: 10.51903/ijgd.v3i2.3991.