

Multi-Agent Cross-Examination for Hallucination Detection and Evidence Localization in Large Language Models

Xu Zhou^{1*}, Michelle Carter², Kai Wang³, Andrew Mitchell⁴

¹Department of Computer Science, University of Massachusetts Amherst, Amherst, MA, USA

²Department of Computer Science, University of Colorado Boulder, Boulder, CO, USA

³Department of Computer Science & Engineering, Washington University in St. Louis, St. Louis, MO, USA

⁴Department of Computer Science, University of Pittsburgh, Pittsburgh, PA, USA

zhou.xu.work@gmail.com

Abstract

Response-level hallucination labels do not identify the decisive text, while repeated judgments add cost without a distinct adversarial role. We evaluate a Detector–Critic–Arbiter framework on 4,507 HaluEval General records. The Detector estimates response risk from lexical and structural views; the Critic scores 27,335 evidence units; and an out-of-fold Arbiter combines probabilities, disagreement, evidence concentration, and structure. A stratified 70/15/15 split, development-tuned thresholds, 300 bootstrap replicates, and paired held-out tests support evaluation. Five-run self-consistency achieved the best macro-F1 (0.642), AUROC (0.721), Brier score (0.211), and accuracy at 80% selective coverage (0.793). Cross-examination achieved the highest AUPRC (0.365), evidence-unit precision (0.667), character precision (0.660), and Hit@1 (0.739), with macro-F1 0.625, ECE 0.290, and character F1 0.444 versus 0.451 for self-consistency. Ablations showed that the learned Arbiter improved ranking over a simple mean, but explicit disagreement features did not improve AUPRC. Capability and refusal errors were easier than semantic or factual errors. Thus, self-consistency is stronger for balanced screening, whereas cross-examination yields more selective, better-ranked evidence for review.

Keywords—Hallucination Detection, Evidence Localization, Multi-Agent Systems, Cross-Examination, Self-Consistency, HaluEval, Calibration

A. INTRODUCTION

Large language models can mix correct content with unsupported claims, mistaken computations, fabricated details, incomplete outputs, and unwarranted refusals. Ji et al. classify these failures as departures from source faithfulness or world knowledge [1]. A closely related operational pattern appears in ambiguity-aware HDFS anomaly detection, where retrieved failure narratives and selective refusal separate actionable evidence from uncertain cases [2]. HaluEval turns the language-model problem into response-level judgments [3], while human-uncertainty distillation shows why ambiguous examples should not be forced into overconfident hard decisions [4]. TruthfulQA further demonstrates that larger models can reproduce common falsehoods [5]. Hallucination detection is therefore a verification task, not a surface-quality test.

Existing methods provide complementary signals. SelfCheckGPT measures sample inconsistency [6], whereas evidence-grounded RAG ties answers to retrieved disclosures [7]. FActScore decomposes generations into atomic facts [8], and long-document RAG preserves clause evidence across

contracts [9]. QAFactEval and QAGS recast consistency as question answering [10], [11], while code retrieval–summary integration follows a find-then-explain workflow [12]. SummaC and FactCC classify entailment or conflict [13], [14]; FEVER supplies claim–evidence supervision [15]; and summarization research shows that fluency can conceal factual divergence [16]. Together, these methods motivate both a response label and localized supporting text.

Localization is essential because a reviewer needs a compact span to verify, correct, or escalate. ERASER formalized rationale evaluation [17], and financial-risk dashboard research shows that visual hierarchy affects explanation usability [18]. Rationalizing predictors select concise text while preserving behavior [19], while progressive PDF rendering prioritizes useful regions before full display [20]. BERT encoders [21] and traffic-sign explanation cards [22] separate representation from presentation. SNLI-style inference resources strengthen verification [23], but one end-to-end score still conflates detection, evidence selection, and adjudication.

Agentic verification separates these roles. Self-consistency aggregates reasoning paths [24], while spectral rank aggregation combines noisy judgments [25]. Multiagent debate and ChatEval use interacting judges [26], [27], and AIOps systems decompose fault analysis into evidence-bearing stages [28]. Self-Refine and Reflexion apply feedback [29], [30], while CI diagnosis links logs to repair recommendations [31]. CRITIC and Chain-of-Verification structure explicit checks [32], [33]; retrieved normal prototypes explain OpenStack anomalies [34]; and DoLa and FacTool add contrastive decoding or tools [35], [36]. These studies support role diversity, but do not establish three-role superiority over repeated detection for both classification and localization.

We test this question with a controlled Detector–Critic–Arbiter design. Sparse probabilistic classifiers keep features and probability traces inspectable on CPU. The Detector scores each query–response pair, the Critic scores atomic units, and the Arbiter combines their signals. Baselines are one hybrid detector and five-head bootstrap self-consistency. Contributions are: full evaluation of 4,507 HaluEval General records under a leakage-controlled split; character- and unit-level localization on 772 records; joint analysis of discrimination, calibration, selective coverage, subgroups, runtime, ablations, and paired tests; and identification of a deployment tradeoff rather than a universal winner. Self-consistency improves balanced detection, whereas cross-examination improves evidence ranking and precision.

Three questions guide evaluation: which consensus strategy yields stronger class-balanced detection; whether a Critic improves evidence precision, coverage, and ordering; and whether gains persist across calibration, compute, length, and error type. Macro-F1 is the primary document metric, AUPRC the ranking metric, and character F1 with Hit@1 the localization metrics. Calibration and selective coverage remain separate because ranking, probability quality, and evidence boundaries can disagree.

We isolate architecture from model scale. Debate studies often change prompts, sampling, model size, and interaction together; here, every comparison uses one sparse representation and one partition. Self-consistency changes only the number of bootstrap Detector heads, whereas cross-examination adds unit-level supervision and a learned Arbiter. Differences therefore reflect repeated judgment versus role specialization within this setting, not a hidden change in language representation.

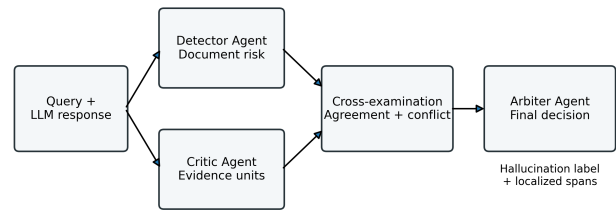


Fig. 1. Detector–Critic–Arbiter Cross-Examination. The Detector Estimates Response Risk, The Critic Identifies Supporting Evidence Units, And The Arbiter Resolves Agreement And Conflict Before Issuing The Label And Spans.

B. METHOD

A. Dataset, Integrity Checks, and Partition

The experiments used the HaluEval General file containing 4,507 newline-delimited JSON records. Each record contains an identifier, a user query, a ChatGPT response, a binary hallucination label, and zero or more annotated hallucination spans. The file includes 815 positive records and 3,692 negative records, producing an 18.08% positive rate. The identifier field contains one duplicated value, so every operation used the immutable row position as the primary key and retained both records. No record was removed from document-level modeling.

Table I. Dataset Partitions And Class Balance

Partition	Records	Hallucination	Non-hallucination	Positive rate
All	4,507	815	3,692	18.08%
Train	3,154	570	2,584	18.07%
Development	676	122	554	18.05%
Test	677	123	554	18.17%

A stratified split assigned 3,154 records to training, 676 to development, and 677 to testing with seed 42. The positive rates remained within 0.13 percentage points across partitions. Vectorizers, scalars, agent parameters, and meta-models were fitted without test records. Development labels were used only to select fixed thresholds and regularization settings. The test partition was opened once for the reported comparisons.

Only the development partition selected operating points. Each method was trained on training data, scored on development, and assigned the macro-F1-maximizing threshold from 0.05 to 0.95. Frozen test thresholds were 0.940 for the single hybrid detector, 0.720 for self-consistency, and 0.620 for cross-examination. The high single-detector value follows from class-balanced weights and its probability scale,

not test adjustment. Structure, word, and character ablations used the same rule.

B. Span Alignment and Evidence Units

The 815 positive records contain 941 annotation strings. Alignment used four deterministic stages: literal matching; Unicode and whitespace normalization; line-wise matching for noncontiguous lists; and longest-common-subsequence matching requiring either 40 normalized characters or a 15-character match covering at least 55% of the annotation. Merged overlaps yielded 887 aligned strings, 940 text matches, and 772 localizable records. The other 43 positives contain only global task-level comments. They remained in detection experiments but were excluded from character localization.

Responses were segmented into evidence units by preserving code, table, list, and line boundaries and splitting long prose lines at sentence boundaries. Short fragments were merged with adjacent text. A unit received a positive label when its character interval overlapped an aligned annotation. The procedure produced 27,335 units, of which 2,170 were positive. Character masks retained the original response offsets, so punctuation, code, and table cells contributed to localization precision and recall rather than disappearing during tokenization.

Table II. Annotation Alignment And Evidence-Unit Statistics

Statistic	Value
Raw hallucination annotations	941
Aligned annotations	887
Annotation alignment rate	94.26%
Hallucinated records with localizable evidence	772
Localizable share among hallucinated records	94.72%
Evidence units	27,335
Positive evidence units	2,170
Mean units per response	6.065
Duplicate ID values	1

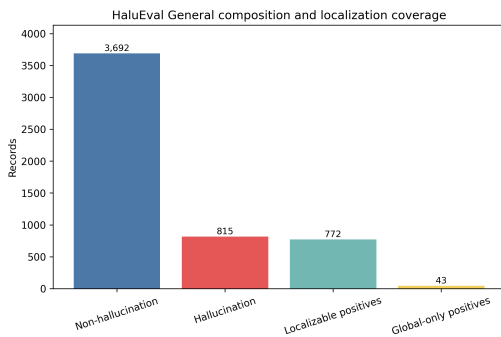


Fig. 2. Class Composition And Localization Coverage. Forty-Three Hallucinated Records Contain Only Global Judgments, While 772 Contain Alignable Response Text.

C. Detector Agent

Detector input concatenates query and response. Its lexical view uses sublinear TF-IDF word unigrams and bigrams (minimum document frequency 2; 8,000 features), and its subword view uses character-within-word 3–5-grams (6,000 features), following the sparse representation of Manning et al. [37]. Twenty structural features cover length, formatting, digits, punctuation, code, URLs, refusal cues, delimiter balance, query overlap, repetition, sentence and line statistics, evidence-unit count, and termination. These lightweight views align with inspectable anomaly pipelines [38] and resource-aware DDoS mitigation [39].

The three views were horizontally concatenated and fitted with a class-balanced linear logistic classifier trained by averaged stochastic gradient descent. The fixed regularization coefficient was 6.67×10^{-6} , training used twenty epochs, and the random seed was 42. Modeling and probability metrics were implemented with scikit-learn [40]. For response i with sparse feature vector x_i , the Detector probability was

$$p_D(i) = \sigma(w_D^T x_i + b_D).$$

The single-agent hybrid baseline used this probability directly. Word-only, character-only, and structure-only detectors measured the value of each feature family. The majority baseline always predicted the negative class.

D. Self-Consistency Baseline

The self-consistency system trained five Detector heads on independent bootstrap samples of the training partition. All heads shared the fitted representation but used separate resampled rows and seeds. The response score was the arithmetic mean of the five probabilities, and the standard deviation quantified disagreement. This design follows the central self-consistency principle of aggregating distinct stochastic paths [24] without adding a new role or evidence objective. It therefore isolates the benefit of repeated judgment from the benefit of cross-examination.

$$p_{SC}(i) = (1/5) \sum_{k=1}^5 p_D^{(k)}(i).$$

E. Critic Agent and Evidence Localization

The Critic operated on each evidence unit together with the query, the preceding unit, and the following unit. Its sparse representation used 7,000 word features, 7,000 character features, and thirteen unit-level structural variables: relative position, character and token length, digit and punctuation ratios, table and code markers, HTML tags, list status, refusal phrases, query overlap, final-unit status, and delimiter imbalance. A class-balanced averaged logistic classifier

produced $p_C(i,j)$, the probability that unit j in response i contained annotated hallucination evidence.

$$p_C(i,j) = \text{sigma}(w_C^T u_{\{i,j\}} + b_C).$$

The Critic threshold was selected on the development partition to maximize micro character F1 after document gating. For a positive document, every unit above the threshold was selected; when no unit crossed the threshold, the highest-scoring unit was retained. For a negative document, the evidence set was empty. Adjacent selected intervals were merged before character scoring. The single-agent localization baseline applied its document detector to individual units, and the self-consistency localization baseline averaged the five unit scores.

F. Arbiter Agent and Cross-Examination Features

The Arbiter combines Detector and Critic signals. Features include Detector, word, and character probabilities; maximum, mean, top-k, dispersion, threshold fraction, and length-weighted Critic coverage; probability gaps; response structure; and unit count. Two-fold out-of-fold predictions create leakage-free training meta-features. A class-balanced logistic model with $C = 0.8$ is then fitted, and refitted base agents generate development and test features. The fusion keeps component contributions testable, as in perception fusion [41], VM residual fusion [42], and uncertainty-calibrated late fusion [43], while treating disagreement as uncertain evidence rather than a verdict [44].

$$p_A(i) = \text{sigma}(w_A^T [p_D(i), g(p_C(i,*)), d(i), s(i)] + b_A),$$

where g aggregates Critic scores, d contains disagreement terms, and s contains structural features. The final document label was one when p_A exceeded the development-selected threshold. The final evidence set was the Critic selection gated by that label. This sequence preserves the cross-examination logic: the Detector makes an accusation, the Critic identifies the textual basis or contradicts the accusation through diffuse low scores, and the Arbiter resolves the record using both views.

G. Metrics, Calibration, and Statistical Tests

Document detection was evaluated with accuracy, balanced accuracy, positive-class precision, recall and F1, macro-F1, Matthews correlation, AUROC, and area under the precision-recall curve. Macro-F1 was the threshold-selection objective because the negative class comprises 81.83% of the data. Calibration was measured with the Brier score [45] and ten-bin expected calibration error following Guo et al. [46]. Selective evaluation ranked records by confidence $\max(p, 1-p)$ and measured accuracy at 80% coverage.

Localization used micro character precision, recall, and F1; macro character F1 across localizable records; mean character intersection-over-union; evidence-unit precision, recall, and F1; and Hit@1, which equals one when the

highest-scoring unit overlaps gold evidence. Confidence intervals used 300 class-stratified bootstrap replicates, following the bootstrap framework of Efron and Tibshirani [47]. Pairwise macro-F1 differences used 500 paired stratified replicates. Exact McNemar tests compared paired correctness outcomes [48]. All modeling and metrics were implemented with scikit-learn [40].

Two additional safeguards controlled optimistic evaluation. First, the Arbiter never received base-agent predictions generated by a model trained on the same training row; its training features came from two-fold cross-fitting. Second, localization thresholds were selected on development character F1 and then applied unchanged to test units. The single-agent, self-consistency, and Critic thresholds were 0.500, 0.170, and 0.260, respectively. The lower consensus thresholds arise because unit probabilities from bootstrap heads and the dedicated Critic occupy different scales. Threshold comparison therefore focuses on the resulting precision and recall, not on the numeric threshold itself.

AUPRC was emphasized alongside AUROC because the positive prevalence is only 18.17% in the test partition. AUROC weights true-positive and false-positive rates across all thresholds, whereas AUPRC directly reflects how precisely hallucinated records are recovered as recall increases. Hit@1 complements character F1 because a reviewer commonly inspects the highest-ranked unit before evaluating the entire marked region. Reporting both measures prevents a system that selects a very large span from receiving an apparently strong hit rate without paying a precision penalty.

Table Iii. Fixed Experimental Configuration

Component	Implementation	Fixed setting
Partition	70/15/15 stratified split	Seed 42; 3,154/676/677 records
Word view	TF-IDF word 1-2 grams	min_df=2; max_features=8,000; sublinear TF
Character view	TF-IDF char_wb 3-5 grams	min_df=2; max_features=6,000
Structural view	20 deterministic features	length, format, overlap, refusal, truncation
Detector Agent	Balanced linear logistic classifier	SGD log-loss; alpha=6.67e-6; 20 epochs; averaged weights
Self-consistency	Five bootstrap detector heads	mean probability; independent seeds
Critic Agent	Unit-level word+character TF-IDF	7,000/7,000 features; SGD alpha=1e-5; 20 epochs
Arbiter Agent	OOF stacked logistic regression	2-fold cross-fitted meta-features; C=0.8
Localization	Character and evidence-unit scoring	unit threshold tuned on development

Component	Implementation	Fixed setting
Uncertainty	10-bin ECE and Brier score	confidence-based 80% selective coverage

C. RESULTS AND DISCUSSION

A. Document-Level Hallucination Detection

Table IV reports the held-out results. The majority classifier reached 0.818 accuracy because of class imbalance but had zero hallucination recall and a macro-F1 of 0.450. Structural cues alone raised macro-F1 to 0.605, confirming that response length, formatting, refusals, and truncation carry real signal. Word and character agents reached 0.618 and 0.623 macro-F1. Combining all views produced 0.629 macro-F1 and 0.707 AUROC, with 70 true positives, 53 false negatives, and 129 false positives.

Table IV. Held-Out Document-Level Detection Results

Method	Accuracy	Precision	Recall	F1	Macro-F1	AUROC	AUPRC
Majority	0.818	0.000	0.000	0.000	0.450	0.500	0.182
Structure only	0.706	0.321	0.553	0.406	0.605	0.693	0.301
Word agent	0.716	0.336	0.577	0.425	0.618	0.703	0.304
Character agent	0.709	0.339	0.634	0.442	0.623	0.705	0.304
Single hybrid	0.731	0.352	0.569	0.435	0.629	0.707	0.310
Self-consistency	0.780	0.403	0.439	0.420	0.642	0.721	0.346
Cross-examination	0.768	0.375	0.415	0.394	0.625	0.707	0.365

Five-run self-consistency delivered the best balanced document performance: macro-F1 0.642, AUROC 0.721, and accuracy 0.780. Its 95% bootstrap interval for macro-F1 was [0.599, 0.684]. Averaging bootstrap heads reduced false positives from 129 to 80, although hallucination recall fell from 0.569 to 0.439. This pattern shows that repeated judgment acted primarily as a variance-reduction mechanism. It removed unstable positive predictions more aggressively than it recovered missed hallucinations.

Cross-examination achieved 0.625 macro-F1 and 0.707 AUROC, so it did not exceed the single hybrid detector or self-consistency. Its document-level advantage was AUPRC: 0.365 versus 0.346 for self-consistency and 0.310 for the single detector. The higher AUPRC indicates better ranking of scarce positives, although the selected threshold favored precision and produced 0.415 recall. The result supports ranking and review rather than universal classification

superiority. Similar distinctions between ranking and one operating point arise in personalized recommendation [49], counterfactual learning-to-rank [50], and data-driven marketing decisions [51].

Ranking and operating-point results differ because the Arbiter precision–recall curve exceeds the single detector across much of low-to-moderate recall, raising AUPRC, while its conservative threshold yields only 51 true positives. Lowering the threshold restores recall and false positives. Self-consistency places its selected point nearer the best macro-F1 balance by averaging extreme Detector scores. Cross-examination therefore contains useful ordering information that its binary threshold does not fully realize.

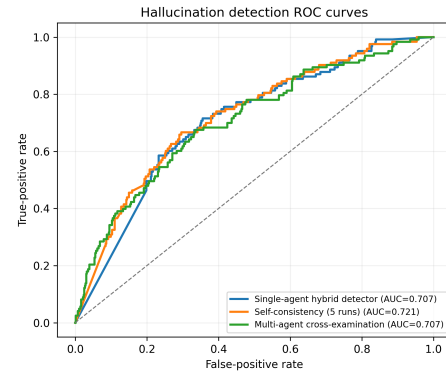


Fig. 3. ROC curves for the three principal systems on the held-out partition. Self-consistency produced the highest AUROC.

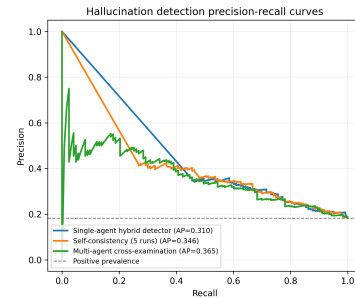


Fig. 4. Precision–Recall Curves. Cross-Examination Achieved The Largest Area Under The Curve, Reflecting Stronger Positive Ranking Under Class Imbalance.

Test-set confusion matrices

	Single-agent hybrid detector		Self-consistency (5 runs)		Multi-agent cross-examination	
Gold	No	Yes	No	Yes	No	Yes
No	425	129	474	80	469	85
Yes	53	70	69	54	72	51
	Predicted No	Predicted Yes	Predicted No	Predicted Yes	Predicted No	Predicted Yes

Fig. 5. Held-Out Confusion Matrices At Development-Selected Thresholds. Self-Consistency Reduced False Positives, While Both Consensus Systems Missed More Positives Than The Single Hybrid Detector.

B. Evidence Localization

Localization was evaluated on the 115 positive test records with alignable character evidence. The single-agent baseline reached character F1 0.443 and unit F1 0.488. Self-consistency achieved the highest character F1, 0.451, and the highest unit F1, 0.493. Cross-examination reached character F1 0.444, nearly matching the two baselines, but its precision-recall profile was distinctly selective. Character precision rose to 0.660 and unit precision to 0.667, while character recall fell to 0.335 and unit recall to 0.333.

Table V. Evidence Localization On 115 Localizable Test Records

Method	Char P	Char R	Char F1	Unit P	Unit R	Unit F1	Hit@1
Single hybrid	0.582	0.358	0.443	0.562	0.432	0.488	0.643
Self-consistency	0.625	0.352	0.451	0.587	0.425	0.493	0.643
Cross-examination	0.660	0.335	0.444	0.667	0.333	0.444	0.739

The most meaningful multi-agent gain was Hit@1: 0.739 versus 0.643 for both baselines. In practical review, the first highlighted unit is often the unit inspected first. Cross-examination therefore placed a gold-overlapping unit at the top for approximately ten additional percentage points of localizable responses. The lower recall arose because the Arbiter suppressed entire evidence sets for false-negative documents and the Critic threshold selected fewer units. This is consistent with the system's higher precision and AUPRC: it concentrated evidence instead of covering every annotated span.

Character precision rewards tight boundaries; Hit@1 asks whether the highest-ranked unit intersects gold evidence. Cross-examination improved both, so the gain was neither indiscriminate span shrinking nor one correct unit surrounded by unrelated text. Its recall loss remains material, making it suitable for evidence-first triage rather than exhaustive annotation. A review interface can open on the top unit and reveal lower-ranked context on demand.

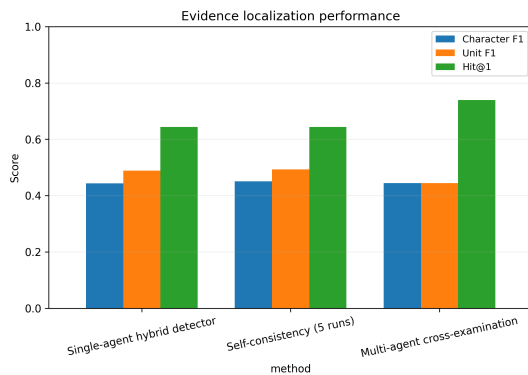


Fig. 6. Character F1, Evidence-Unit F1, And Hit@1. Cross-Examination Improved Top-Ranked Evidence While Preserving Character F1 Close To The Baselines.

C. Ablation Analysis

Table VI separates the contribution of aggregation, Critic evidence, disagreement, and structural variables. A simple mean of Detector and Critic scores produced 0.631 macro-F1 and 0.324 AUPRC. The learned Arbiter raised AUPRC to 0.365, demonstrating that nonlinear operating behavior emerged from the stacked feature combination rather than from averaging alone. Removing the Critic retained 0.353 AUPRC, so evidence features contributed 0.012 absolute AUPRC to the full model.

Table Vi. Detector-Critic-Arbiter Ablations

Configuration	Precision	Recall	Macro-F1	AUROC	AUPRC
Detector + Critic mean	0.369	0.480	0.631	0.700	0.324
Arbiter without critic	0.388	0.382	0.625	0.716	0.353
Arbiter without disagreement	0.383	0.398	0.626	0.709	0.367
Arbiter without structural features	0.364	0.455	0.626	0.692	0.357
Full Detector-Critic-Arbiter	0.375	0.415	0.625	0.707	0.365

Ablations show that not every cross-examination feature helped. Removing explicit disagreement variables increased AUPRC slightly from 0.365 to 0.367 with unchanged macro-F1, indicating redundancy with base probabilities. Removing structural features reduced AUROC from 0.707 to 0.692 but raised recall from 0.415 to 0.455; format cues improved ranking while making positive decisions more conservative. The full design was retained for its strongest localization trace and role separation, not because each scalar feature improved every document metric. This interpretation is consistent with image-detection ablations that separate feature-extractor changes from attention effects [52].

D. Calibration and Selective Prediction

Calibration results reinforce the difference between consensus and cross-examination. Self-consistency achieved the lowest Brier score, 0.211, and the lowest expected calibration error, 0.202. At 80% selective coverage its accuracy was 0.793, compared with 0.738 for the single hybrid detector and 0.753 for cross-examination. Probability averaging therefore improved both discrimination and confidence reliability.

Table Vii. Calibration And Confidence-Based Selective Results

Method	Brier	ECE	Coverage at conf. ≥ 0.80	Accuracy at conf. ≥ 0.80	Accuracy at 80% coverage
Single hybrid	0.270	0.272	97.6%	0.726	0.738
Self-consistency	0.211	0.202	77.0%	0.791	0.793
Cross-examination	0.222	0.290	7.5%	0.529	0.753

The Arbiter was poorly calibrated: ECE was 0.290, only 7.5% of records reached confidence 0.80, and accuracy within that subset was 0.529. Its reliability curve shows underconfidence in several middle bins and a sharp highest-bin change. Improved positive ranking therefore did not produce dependable absolute confidence. Post-hoc calibration on a representative stream is required before confidence controls acceptance or escalation. The same requirement appears in credit-risk management [53], AI-assisted financial risk analysis [54], risk-aware auto-bidding [55], uncertainty-macro fusion for VIX direction [56], and constrained budgeting [57].

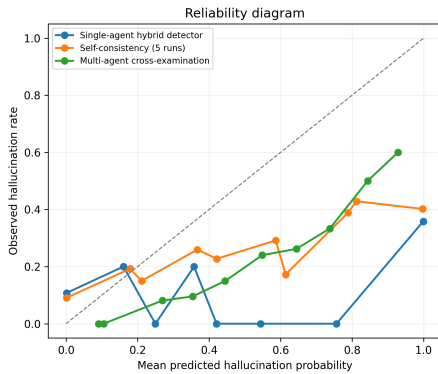


Fig. 7. Reliability Curves. Five-Run Self-Consistency Was The Best Calibrated System; The Arbiter Probabilities Require Separate Calibration Before Confidence-Gated Use.

E. Response Length and Error Categories

Length stratification exposes where each method changed behavior. Self-consistency led the shortest quartile with macro-F1 0.693 and the second quartile with 0.638. Cross-examination matched the second quartile at 0.637 and led the third quartile at 0.547, where the single detector and self-consistency scored 0.515 and 0.530. All methods weakened on longer responses because positive evidence occupied a smaller fraction of the text and the sparse features mixed correct and incorrect units. In the longest quartile, the single detector recovered the strongest macro-F1, 0.634, because its higher positive recall outweighed additional false positives.

Table Viii. Detection By Response-Length Quartile

Method	Length group	n	Mean words	Macro-F1	Hallucination recall
Single hybrid	Q1 shortest	172	36.3	0.683	0.680
Single hybrid	Q2	172	70.6	0.615	0.481
Single hybrid	Q3	171	99.2	0.515	0.375
Single hybrid	Q4 longest	162	129.3	0.634	0.567
Self-consistency	Q1 shortest	172	36.3	0.693	0.580
Self-consistency	Q2	172	70.6	0.638	0.370
Self-consistency	Q3	171	99.2	0.530	0.188
Self-consistency	Q4 longest	162	129.3	0.600	0.400
Cross-examination	Q1 shortest	172	36.3	0.662	0.520
Cross-examination	Q2	172	70.6	0.637	0.407
Cross-examination	Q3	171	99.2	0.547	0.250
Cross-examination	Q4 longest	162	129.3	0.576	0.333

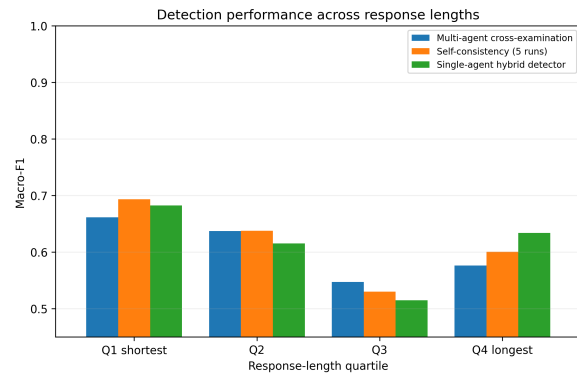


Fig. 8. Macro-F1 Across Response-Length Quartiles. The Third Quartile Was The Only Length Group Led By Cross-Examination.

Error-Category Recall Exposes The Main Limitation. The single detector recalled 91.9% of capability/refusal positives and 76.9% of numeric, code, or table positives, but 34.7% of semantic or factual positives; cross-examination recalled 81.1%, 46.2%, and 19.4%. Refusal and formatting failures create recurring lexical patterns, whereas semantic errors require external knowledge, arithmetic, or task-specific reasoning. The gap would be more severe in specialized domains such as cancer image classification [58] or MRD prediction for multiple myeloma [59], whose evidence is absent from HaluEval. The Critic improves ranking within

the difficult set, but its conservative threshold rejects many cases without recurring surface cues.

Table IX. Positive-Class Recall By Error Category

Method	Category	Positive records	Recall	Mean probability
Single hybrid	Capability/refusal	37	0.919	0.920
Single hybrid	Incomplete/truncated	1	1.000	1.000
Single hybrid	Numeric/code/table	13	0.769	0.840
Single hybrid	Semantic/factual	72	0.347	0.369
Self-consistency	Capability/refusal	37	0.757	0.840
Self-consistency	Incomplete/truncated	1	1.000	0.800
Self-consistency	Numeric/code/table	13	0.615	0.664
Self-consistency	Semantic/factual	72	0.236	0.349
Cross-examination	Capability/refusal	37	0.811	0.759
Cross-examination	Incomplete/truncated	1	1.000	0.695
Cross-examination	Numeric/code/table	13	0.462	0.623
Cross-examination	Semantic/factual	72	0.194	0.480

F. Efficiency and Paired Statistical Tests

Efficiency measurements isolate scoring on precomputed sparse features. For the 677-record test batch, the single detector required one decision and 0.482 ms; five-head self-consistency required five decisions and 2.452 ms; and cross-examination required three decisions and 2.549 ms because the Critic also scored 4,017 evidence units. Serialized models remained below one megabyte. Both consensus strategies were computationally light, and cross-examination used 40% fewer agent decisions than self-consistency. Selective offloading and compact proxies are also central to hybrid-cloud LLM deployment [60] and edge-deployable video-quality prediction [61].

Table X. Model-Scoring Efficiency On The Held-Out Batch

Method	Agent decisions	Batch ms	ms/response	Model MB	Macro-F1
Single hybrid	1	0.482	0.0007	0.279	0.629
Self-consistency	5	2.452	0.0036	0.867	0.642
Cross-examination	3	2.549	0.0038	0.560	0.625

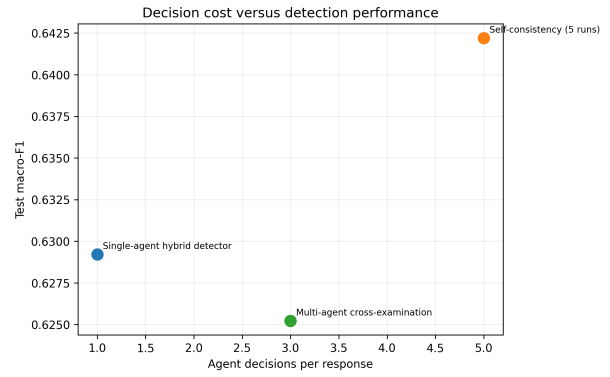


Fig. 9. Agent decisions per response versus held-out macro-F1. Self-consistency obtained the highest macro-F1 at five decisions; cross-examination used three decisions.

Paired bootstrap tests found no macro-F1 improvement for cross-examination. Its difference from the single hybrid detector was -0.002 with a 95% interval of [-0.035, 0.032] and $p = 0.888$. Its difference from self-consistency was -0.016 with interval [-0.049, 0.018] and $p = 0.368$. Exact McNemar testing told a different story for raw correctness against the single detector: cross-examination corrected 54 records missed by the baseline, while the baseline corrected 29 records missed by cross-examination, yielding $p = 0.008$. Most of those corrections were negative records, so accuracy improved while positive recall declined. Against self-consistency, McNemar $p = 0.341$ indicated no paired correctness difference.

Table XI. Paired Significance Tests For Cross-Examination

Comparison	Δ Macro-F1	95% CI	Bootstrap p	McNemar p	Cross-exam. only correct	Baseline only correct
Cross-exam. vs Single-agent hybrid detector	-0.002	[-0.035, 0.032]	0.888	0.008	54	29
Cross-exam. vs Self-consistency (5 runs)	-0.016	[-0.049, 0.018]	0.368	0.341	23	31

G. Held-Out Cross-Examination Trace

Figure 10 shows HaluEval record 120 from the test partition. The user requested an acoustic guitar riff. The response began with an unwarranted capability refusal and then supplied five useful construction steps. The gold annotation marks only the first unit. The Detector assigned probability 1.000, the Critic assigned 1.000 to the refusal and values below 0.001 to the remaining units, and the Arbiter assigned 0.697, above its 0.620 threshold. The final evidence set exactly matched the annotated unit. This trace demonstrates the intended

behavior: the final label does not require highlighting the entire response, and correct or useful material remains outside the evidence boundary.

Held-out cross-examination trace (HaluEval ID 120)

Query: Create an acoustic guitar riff.

Detector probability = 1.000 | Arbiter probability = 0.697 | Final decision = hallucination

Gold evidence and Critic selected evidence coincide in Unit 1; the remaining response units are retained as non-evidence.

Unit 1	Critic p=1.000	gold=yes	selected=yes
Unfortunately, as an AI language model, I cannot create an acoustic guitar riff in the traditional sense, as I don't have access to an instrument.			
Unit 2	Critic p=0.000	gold=no	selected=no
However, I can provide you with a general structure for creating a guitar riff, which you can use as inspiration for your own:			
Unit 3	Critic p=0.000	gold=no	selected=no
1) Start with a basic chord progression, such as G-C-D or Am-F-C-G.			
Unit 4	Critic p=0.000	gold=no	selected=no
2) Experiment with different drumming patterns and finger-picking techniques to create a unique sound.			
Unit 5	Critic p=0.000	gold=no	selected=no
3) Incorporate hammer-ons, pull-offs, and slides to add texture and variety to your riff.			
Unit 6	Critic p=0.000	gold=no	selected=no
4) Experiment with different tempos and dynamics to create a sense of energy and emotion.			
Unit 7	Critic p=0.000	gold=no	selected=no
5) Don't be afraid to try new things and take risks in your playing - the beauty of creating a guitar riff is that there are no hard and fast rules, so let your creativity run wild!			

The probabilities correspond to the Detector, Critic, and Arbiter decisions for this held-out record.

Fig. 10. Detector–Critic–Arbiter decisions for held-out HaluEval record 120. The Critic isolates the annotated refusal while leaving the useful instructions unselected.

D. LIMITATIONS

The evaluation is in-domain on HaluEval General. A random stratified split measures discrimination within one benchmark distribution; it does not establish transfer to newer model families, retrieval-augmented systems, multilingual dialogue, or scientific and medical domains. Language context can alter learning and engagement, as bilingual computer-science education studies demonstrate [62]. Cross-domain validity also depends on shared-feature assumptions [63] and observation design, as illustrated by digital-twin dispatching [64]. The reported thresholds therefore require cross-benchmark validation before external use.

The agents are sparse probabilistic classifiers rather than generative LLM agents. This isolates architecture and keeps every score reproducible on CPU, but omits external retrieval, symbolic arithmetic, source attribution, and open-ended questioning. Evidence-grounded RAG [7], long-document clause retrieval [9], code retrieval–summary integration [12], OpenStack prototype retrieval [34], and CI repair pipelines [31] operate in a richer and more expensive regime. The low recall on semantic and factual errors follows directly from this difference; the present results do not quantify that regime.

Localization covers 772 of 815 positive records because 43 records contain only global annotation comments. The alignment algorithm recovered 94.26% of annotation strings, including normalized and high-overlap matches. Character metrics exclude records without literal evidence and therefore measure localizable hallucinations rather than every form of task failure. The unit segmentation also treats lines, code, lists, and sentences as atomic boundaries; a different segmentation changes unit precision and recall even when character masks remain fixed.

The multi-agent system is not the best document classifier in this evaluation. Self-consistency has higher macro-F1,

AUROC, Brier score, and selective accuracy. Cross-examination has higher AUPRC, character precision, unit precision, and Hit@1, but lower recall and poor probability calibration. The ablation results also show that explicit disagreement variables are redundant. These outcomes limit the valid claim to a precision-oriented localization and ranking advantage.

Runtime measures scoring after feature construction; end-to-end latency also includes parsing, segmentation, TF-IDF transformation, and deployment. Error categories are deterministic diagnostic groups rather than independent human labels, and the incomplete group contains one positive test record. Category recall therefore supports qualitative diagnosis, not a population estimate. Production use would also require drift monitoring, safe retraining, and rollback controls of the kind formalized by DriftGuard [65].

E. CONCLUSION

A full evaluation on 4,507 HaluEval General records compared single-agent detection, five-run self-consistency, and Detector–Critic–Arbiter cross-examination under one data split and one evidence-alignment protocol. Self-consistency produced the strongest balanced document detection, with macro-F1 0.642, AUROC 0.721, Brier score 0.211, and 0.793 accuracy at 80% selective coverage. Cross-examination produced the strongest positive ranking and the most selective evidence, with AUPRC 0.365, character precision 0.660, unit precision 0.667, and Hit@1 0.739. Its character F1 of 0.444 remained close to the self-consistency value of 0.451, but its document macro-F1 of 0.625 did not improve on the baselines.

The results define two distinct deployment choices. When the primary objective is balanced response-level screening with reliable confidence, five-run self-consistency is the stronger configuration. When the primary objective is to rank rare hallucinations and present a compact first piece of evidence to a reviewer, Detector–Critic–Arbiter cross-examination is the stronger configuration. The remaining performance gap lies in semantic verification: every system detects recurring capability and formatting failures more readily than unsupported factual content. Adding grounded retrieval and domain tools to the Critic is the direct next architectural step, while preserving the separate Arbiter so that evidence quality and document risk remain independently testable.

F. REFERENCES

- [1] Z. Ji et al., “Survey of hallucination in natural language generation,” *ACM Comput. Surv.*, vol. 55, no. 12, Art. no. 248, pp. 1–38, 2023.
- [2] 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.
- [3] J. Li, X. Cheng, W. X. Zhao, J.-Y. Nie, and J.-R. Wen, “HaluEval: A large-scale hallucination evaluation benchmark for large language models,” in Proc. 2023 Conf. Empirical Methods in Natural Language Processing, 2023, pp. 6449–6464.
- [4] 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.
- [5] S. Lin, J. Hilton, and O. Evans, “TruthfulQA: Measuring how models mimic human falsehoods,” in Proc. 60th Annu. Meeting Assoc. Comput. Linguistics, 2022, pp. 3214–3252.
- [6] P. Manakul, A. Liusie, and M. J. F. Gales, “SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models,” in Proc. 2023 Conf. Empirical Methods in Natural Language Processing, 2023, pp. 9004–9017.
- [7] 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.
- [8] S. Min et al., “FactScore: Fine-grained atomic evaluation of factual precision in long form text generation,” in Proc. 2023 Conf. Empirical Methods in Natural Language Processing, 2023, pp. 12076–12100.
- [9] 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.
- [10] R. Fabbri, C.-S. Wu, W. Liu, and C. Xiong, “QAFactEval: Improved QA-based factual consistency evaluation for summarization,” in Proc. 2022 Conf. North American Chapter Assoc. Comput. Linguistics, 2022, pp. 2587–2601.
- [11] Wang, K. Cho, and M. Lewis, “Asking and answering questions to evaluate the factual consistency of summaries,” in Proc. 58th Annu. Meeting Assoc. Comput. Linguistics, 2020, pp. 5008–5020.
- [12] 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.
- [13] P. Laban, T. Schnabel, P. N. Bennett, and M. A. Hearst, “SummaC: Re-visiting NLI-based models for inconsistency detection in summarization,” Trans. Assoc. Comput. Linguistics, vol. 10, pp. 163–177, 2022.
- [14] W. Kryściński, B. McCann, C. Xiong, and R. Socher, “Evaluating the factual consistency of abstractive text summarization,” in Proc. 2020 Conf. Empirical Methods in Natural Language Processing, 2020, pp. 9332–9346.
- [15] J. Thorne, A. Vlachos, C. Christodoulopoulos, and A. Mittal, “FEVER: A large-scale dataset for fact extraction and verification,” in Proc. 2018 Conf. North American Chapter Assoc. Comput. Linguistics, 2018, pp. 809–819.
- [16] J. Maynez, S. Narayan, B. Bohnet, and R. McDonald, “On faithfulness and factuality in abstractive summarization,” in Proc. 58th Annu. Meeting Assoc. Comput. Linguistics, 2020, pp. 1906–1919.
- [17] J. DeYoung et al., “ERASER: A benchmark to evaluate rationalized NLP models,” in Proc. 58th Annu. Meeting Assoc. Comput. Linguistics, 2020, pp. 4443–4458.
- [18] 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.
- [19] T. Lei, R. Barzilay, and T. Jaakkola, “Rationalizing neural predictions,” in Proc. 2016 Conf. Empirical Methods in Natural Language Processing, 2016, pp. 107–117.
- [20] 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.
- [21] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in Proc. 2019 Conf. North American Chapter Assoc. Comput. Linguistics, 2019, pp. 4171–4186.
- [22] 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.
- [23] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning, “A large annotated corpus for learning natural language inference,” in Proc. 2015 Conf. Empirical Methods in Natural Language Processing, 2015, pp. 632–642.
- [24] X. Wang et al., “Self-consistency improves chain of thought reasoning in language models,” in Proc. Int. Conf. Learning Representations, 2023.
- [25] Z. S. Zhong and S. Ling, “Improved theoretical guarantee for rank aggregation via spectral method,” Inf. Inference, vol. 13, no. 3, Art. no. iaac020, Sep. 2024, doi: 10.1093/imaia/iaac020.

- [26] 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.
- [27] C.-M. Chan et al., "ChatEval: Towards better LLM-based evaluators through multi-agent debate," arXiv:2308.07201, 2023.
- [28] 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.
- [29] Madaan et al., "Self-Refine: Iterative refinement with self-feedback," in *Adv. Neural Inf. Process. Syst.*, vol. 36, 2023, pp. 46534–46594.
- [30] N. Shinn, F. Cassano, A. Gopinath, K. Narasimhan, and S. Yao, "Reflexion: Language agents with verbal reinforcement learning," in *Adv. Neural Inf. Process. Syst.*, vol. 36, 2023, pp. 8634–8652.
- [31] 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.
- [32] Z. Gou et al., "CRITIC: Large language models can self-correct with tool-interactive critiquing," arXiv:2305.11738, 2023.
- [33] S. Dhuliawala et al., "Chain-of-verification reduces hallucination in large language models," arXiv:2309.11495, 2023.
- [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] Y.-S. Chuang et al., "DoLa: Decoding by contrasting layers improves factuality in large language models," arXiv:2309.03883, 2023.
- [36] I.-C. Chern et al., "FacTool: Factuality detection in generative AI—A tool augmented framework for multi-task and multi-domain scenarios," arXiv:2307.13528, 2023.
- [37] D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge, U.K.: Cambridge Univ. Press, 2008.
- [38] Q. Xin, Z. Xu, L. Guo, F. Zhao, and B. Wu, "IoT traffic classification and anomaly detection method based on deep autoencoders," in *Proc. 6th Int. Conf. Computing and Data Science*, 2024, pp. 64–70, doi: 10.54254/2755-2721/69/20241511.
- [39] Wang, Y. He, Z. Shui, Q. Xin, and H. Lei, "Predictive optimization of DDoS attack mitigation in distributed systems using machine learning," in *Proc. 6th Int. Conf. Computing and Data Science*, 2024, pp. 94–99, doi: 10.54254/2755-2721/64/20241350.
- [40] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, no. 85, pp. 2825–2830, 2011.
- [41] 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.
- [42] 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.
- [43] Q. Xin, "Uncertainty-aware late fusion for 3D perception: Confidence calibration and fusion-rule learning," *J. Technol. Informatics Eng.*, vol. 4, no. 1, pp. 215–238, Apr. 2025, doi: 10.51903/jtie.v4i1.485.
- [44] Z. S. Zhong and S. Ling, "Uncertainty quantification of spectral estimator and MLE for orthogonal group synchronization," arXiv:2408.05944, Aug. 2024.
- [45] G. W. Brier, "Verification of forecasts expressed in terms of probability," *Monthly Weather Rev.*, vol. 78, no. 1, pp. 1–3, 1950.
- [46] 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.
- [47] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. New York, NY, USA: Chapman & Hall, 1993.
- [48] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," *Psychometrika*, vol. 12, no. 2, pp. 153–157, 1947.
- [49] Xu, H. Zhou, H. Zheng, M. Zhu, and Q. Xin, "Intelligent classification and personalized recommendation of e-commerce products based on machine learning," in *Proc. 6th Int. Conf. Computing and Data Science*, 2024, pp. 143–149, doi: 10.54254/2755-2721/64/20241365.
- [50] 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, 2025.
- [51] 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.
- [52] Z. Ling, Q. Xin, Y. Lin, G. Su, and Z. Shui, "Optimization of autonomous driving image detection based on RFACnv and triplet attention," in *Proc. 2nd Int. Conf. Software Engineering and Machine Learning*, 2024, pp. 210–217, doi: 10.54254/2755-2721/77/2024MA0067.
- [53] 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, doi: 10.5281/zenodo.14032041.

- [54] 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.
- [55] 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.
- [56] H. Zhou and K. Zhang, "News-based uncertainty and macro-market fusion for VIX direction forecasting: Evidence from a 2015–2024 FRED panel," *J. Technol. Informatics Eng.*, vol. 4, no. 2, pp. 487–501, Aug. 2025, doi: 10.51903/jtie.v4i2.540.
- [57] 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.
- [58] 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, 2020, doi: 10.1088/1742-6596/1651/1/012143.
- [59] J. Chen, J. Xiong, Y. Wang, Q. Xin, and H. Zhou, "Implementation of an AI-based MRD evaluation and prediction model for multiple myeloma," *Front. Comput. Intell. Syst.*, vol. 6, no. 3, pp. 127–131, 2024, doi: 10.54097/zJ4MnbWW.
- [60] 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.
- [61] 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.
- [62] 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.*, 2019, pp. 1–10, doi: 10.1145/3364510.3364518.
- [63] 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, 2025, pp. 559–567.
- [64] 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.
- [65] 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.