

Calibration-Aware Truthful LLMs with Selective Answering and Abstention

Kimberly Green^{1*}, Lei Sun², Robert Baker³, Jing Chen⁴, Justin Nelson⁵

¹Department of Computer Science and Engineering, University of California, Riverside, CA, USA

²Department of Computer Science and Engineering, The Pennsylvania State University, University Park, PA, USA

³Department of Computer Science, Rutgers, The State University of New Jersey, New Brunswick, NJ, USA

⁴Department of Computer Science, University of California, Santa Barbara, CA, USA

⁵Department of Computer Science, Boston University, Boston, MA, USA

kim.green.dev@yahoo.com

Abstract

Large language models can select a fluent falsehood even when their answer score appears decisive, so forced-answer accuracy alone does not determine when an answer should be trusted. This study introduces CASA-Truth, a calibration-aware selective-answering layer that separates candidate selection from confidence estimation and uses the calibrated probability of correctness to answer or abstain. The evaluation used all 817 TruthfulQA questions, 38 categories, and 4,187 MC1-style candidate answers. A question-level five-fold protocol assigned three folds to scorer training, one fold to calibration and threshold selection, and one fold to testing, preventing answer candidates from the same question from crossing data partitions. The fold-tuned candidate scorer attained 72.22% accuracy. Relative to raw softmax confidence, CASA-Truth reduced the Brier Score from 0.2007 to 0.1433, reduced expected calibration error from 0.1826 to 0.0312, and reduced area under the risk-coverage curve from 0.1176 to 0.0954. At 70.01% coverage, selective accuracy reached 85.49%; at 50.06% coverage, it reached 93.40%. The system retained 60.47% coverage while sustaining at least 90% empirical accuracy. Paired bootstrap analysis placed the Brier reduction at 0.0574 with a 95% interval of [0.0427, 0.0711]. These findings show that calibrated abstention converts a fixed truthfulness scorer into a controllable answer policy with substantially better reliability and selective risk.

Keywords: Large Language Models; Hallucination; Truthfulqa; Calibration; Selective Prediction; Abstention; Uncertainty Estimation; Brier Score; Expected Calibration Error

A. INTRODUCTION

Large language models deliver fluent question answering at scale, but fluency is not truthfulness. Few-shot scaling can reproduce training-data misconceptions [1], and instruction tuning improves helpfulness without guaranteeing factual validity [2]. Hybrid cloud-edge designs reduce deployment cost [3], not epistemic error. TruthfulQA targets this gap with questions whose familiar answer may be false [4]. Reliable deployment therefore requires both an answer and an estimate that it is correct.

Calibration gives confidence an empirical meaning: predictions reported at 0.80 should be correct about 80% of the time. The Brier Score measures probabilistic error [5], and its decomposition separates reliability, resolution, and uncertainty [6]. Neural networks often need post-hoc correction [7]. Human-uncertainty distillation uses judgment distributions as calibration supervision [8], while late fusion inserts calibrated confidence into sensor decisions [9]. Platt and isotonic mappings [10], [11], [12], Bayesian binning [13], and beta calibration [14] provide complementary score-to-probability transformations.

In high-cost applications, confidence directly affects action. Medical classifiers [15], [16], autonomous-driving perception systems [17], [18], and financial-risk models [19], [20] all convert scores into consequential decisions. This motivates treating confidence as a decision variable rather than a display value.

Selective prediction adds an answer-or-abstain decision. Coverage is the fraction answered; selective risk is error among accepted cases. Deep selective classification formalized this tradeoff [21], ambiguity-aware log analysis paired evidence retrieval with refusal [22], and SelectiveNet learned rejection jointly [23]. Foundational work established noise-free selection [24] and optimal error-rejection tradeoffs [25]; risk-aware bidding applies analogous constraints to budgeted actions [26]. For factual QA, confidence must rank truths above falsehoods and thresholds must be selected without test leakage.

Maximum softmax probability is a practical error signal [27], while Bayesian dropout [28], deep ensembles [29], and shift studies [30] address broader uncertainty. AIOps analysis [31] and DriftGuard [32] connect uncertainty or drift to diagnosis

and rollback; VM residual fusion separates degradation from shared load [33]. Spectral-estimator analysis provides a theoretical uncertainty-quantification perspective [34]. In NLP, pretrained transformers require explicit calibration [35], [36].

Language models can verbalize uncertainty [37], self-evaluate answers [38], aggregate meaning-equivalent generations [39], or use sampling consistency [40]. Retrieval adds external evidence through financial-disclosure RAG [41], code retrieval-summary integration [42], long-document clause analysis [43], OpenStack prototype retrieval [44], and CI log-to-patch diagnosis [45]. These methods improve evidence but do not calibrate the final answer decision by themselves.

Other pipelines also embed predictions in decisions: recommendation and marketing [46], [47], VIX forecasting and budgeting [48], [49], mobility dispatch [50], DDoS mitigation [51], and IoT anomaly detection [52]. Bilingual CS education further shows that evidence communication depends on user context [53]. CASA-Truth therefore separates confidence estimation from the downstream threshold.

This study focuses on that post-generation decision problem. CASA-Truth receives a question and candidate answers, scores their truthfulness, converts the score pattern and answer-level signals into a calibrated probability of correctness, and applies an abstention rule. The controlled evaluation uses the best answer and each annotated false answer in TruthfulQA as the candidate set. This design removes generation randomness and isolates three questions: which candidate-scoring features improve truthfulness selection, which calibration method produces the most faithful confidence, and how much accuracy can be gained by reducing coverage.

The experimental protocol makes four contributions. First, it evaluates all questions under a question-level five-fold design, so no candidate answer from a test question appears in training or calibration. Second, it compares answer-only word features, answer word-plus-character features, question-answer features, a hybrid scorer, and a fold-tuned ensemble. Third, it compares raw softmax confidence with temperature, Platt, isotonic, beta, and feature-aware calibration using Brier Score, negative log-likelihood, ECE, adaptive ECE, and risk-coverage metrics. Fourth, it selects answer-or-abstain thresholds under explicit deferral costs and reports paired bootstrap intervals for the main improvements.

Candidate accuracy, calibration, and selective utility are distinct. The first asks whether the best answer was selected; the second whether confidence matches observed correctness; the third whether confidence improves accuracy as coverage falls. Spectral rank aggregation likewise separates ordering quality from estimator uncertainty [54]. The experiments report all three because improvement in one need not imply improvement in another.

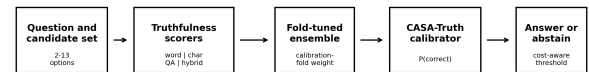
B. METHOD

Dataset and candidate construction. The TruthfulQA release contains 817 English questions distributed across 38 categories, including misconceptions, law, health, sociology, conspiracies, indexical errors, and common confusions [4]. Each record contains a best answer, a set of accepted correct answers, and a set of incorrect answers designed to reflect common falsehoods. The primary experiment follows an MC1-style decision task: the candidate set for each question contains the best answer and every annotated incorrect answer. This construction yields 4,187 candidate answers, with a mean of 5.125 candidates per question and a range from 2 to 13. Candidate order was shuffled with seed 42 separately for every question, eliminating any position convention.

The positive label equals one only for the best answer. The scorer receives each question-answer pair independently and returns a real-valued truthfulness score. The selected answer is the candidate with the largest score. Let $s(q, a_j)$ denote the score for candidate j of question q . The forced-answer prediction is $a^* = \arg\max_j s(q, a_j)$. A prediction is correct when a^* is the best answer. Raw confidence is obtained by applying a within-question softmax to the candidate scores.

Table I
Truthfulqa Data Used In The Experimental Evaluation.

Statistic	Value
Questions	817
Categories	38
Adversarial questions	437
Non-adversarial questions	380
Primary MC1 candidates	4187
Mean candidates per question	5.125
Median candidates per question	5.0
Minimum/maximum candidates	2/13
Mean annotated correct answers	3.475
Mean annotated incorrect answers	4.125
Missing source values	2



Scorer training, confidence calibration, threshold selection, and final testing use disjoint question sets in every outer fold.

Fig. 1. CASA-Truth candidate scoring, calibration, and answer-or-abstain pipeline.

Candidate scorers. The answer-word scorer uses TF-IDF word unigrams and bigrams from the candidate answer. The answer-word-character scorer adds character n-grams, which capture morphology, negation fragments, and lexical patterns that remain useful when word overlap is sparse. The question-answer scorer applies word and character TF-IDF to the concatenated text 'Question: q Answer: a'. The hybrid scorer combines answer-only and question-answer representations with twelve standardized numeric features. TF-IDF follows the standard term-weighting principles described by Salton and Buckley [55]. All word vectorizers used $\min_df=2$, sublinear

term frequency, and at most 30,000 features, except the question-answer word vectorizer, which used 40,000. Character vectorizers used char_wb 3-5 grams, min_df=2, and 30,000 features.

The numeric features were log question length, log answer length, answer-to-question length ratio, token-set Jaccard overlap, rates of negation, hedging, and absolute terms, counts of semicolons, question marks, and commas, an indicator for explicit uncertainty phrases, and log candidate count. Each scorer used L2-regularized logistic regression with C=1.0, balanced class weights, the liblinear solver, and a maximum of 500 iterations. The implementation used scikit-learn [56].

Fold-tuned score ensemble. The strongest question-answer scorer and the hybrid scorer make different errors: the former emphasizes pair semantics, while the latter uses answer form and structural signals. For each outer fold, the ensemble score was $w s_{QA} + (1-w) s_H$. The weight w was selected on the calibration questions from the grid {0.000, 0.125, 0.250, 0.375, 0.500, 0.625, 0.750, 0.875, 1.000}. Selection maximized calibration-set MC1 accuracy; ties were broken by lower Brier Score and then lower negative log-likelihood. The chosen weight was applied unchanged to the outer test questions.

Question-level evaluation split. A stratified five-fold split preserved the adversarial versus non-adversarial proportion. For outer fold f , fold f served as test data, fold $(f+1) \bmod 5$ served as calibration data, and the remaining three folds trained the candidate scorers. Consequently, approximately 60% of questions trained the scorer, 20% fitted calibration maps and thresholds, and 20% remained untouched for testing. Every question appeared exactly once in an outer test fold and once in a calibration fold. All candidate answers belonging to a question remained together, which prevented question-level leakage. The random seed was 42.

Table Ii

Fixed experimental protocol and hyperparameters.

Component	Setting
Random seed	42
Outer evaluation	5 question-level stratified folds
Per-fold allocation	3 train / 1 calibration / 1 test folds
Stratification variable	Adversarial vs non-adversarial type
Word features	TF-IDF unigrams and bigrams; min_df=2
Character features	char_wb TF-IDF, 3-5 grams; min_df=2
Classifier	L2 logistic regression, C=1.0, balanced classes
Calibration binning	10 equal-width and 10 equal-mass bins
Ensemble weights	9-point grid from 0.000 to 1.000
Bootstrap	2,000 paired question-level resamples
Primary metrics	MC1 accuracy, ECE, Brier, NLL, AURC, E-AURC

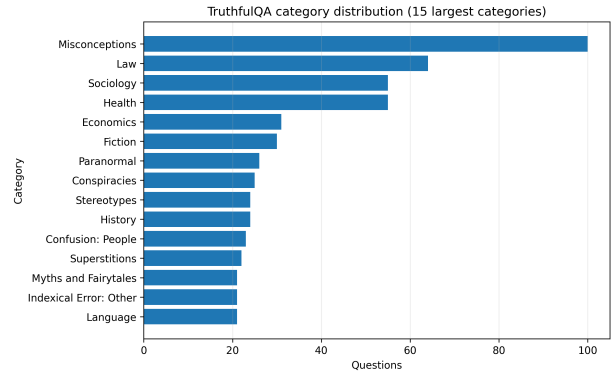


Fig. 2. Distribution of the fifteen largest TruthfulQA categories.

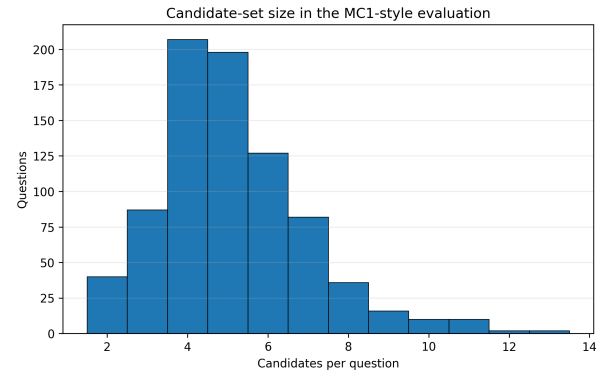


Fig. 3. Number of candidate answers per question in the primary evaluation.

Calibration baselines. Temperature scaling fitted one positive temperature T by minimizing multiclass negative log-likelihood on calibration questions [7]. It transformed each candidate score to s/T and preserved the selected answer. Platt scaling fitted logistic regression to the log-odds of the raw top-answer probability [12]. Isotonic regression fitted a nondecreasing piecewise-constant mapping, following the use of isotonic calibration in probabilistic classification [10], [11]. Beta calibration fitted logistic regression to $\log(p)$ and $\log(1-p)$, providing a richer parametric family than a single sigmoid [14]. Each fitted map was applied only to the corresponding outer test fold.

CASA-Truth calibration. The proposed calibrator predicts whether the selected answer is correct. Its inputs are raw top probability, second probability, probability margin, score margin, normalized entropy, top score, score standard deviation, candidate count, question length, answer length, length ratio, question-answer overlap, negation rate, hedge rate, absolute-term rate, an explicit-uncertainty indicator, and the adversarial-type indicator. Count features were log transformed. A StandardScaler and logistic regression with C=0.5 were fitted on each calibration fold. The model returned the confidence used for reliability evaluation and selective ordering. Three ablations progressively added structural, lexical, and type features to the score-only set.

Calibration metrics. For a binary correctness label y_i and reported confidence p_i , the Brier Score was the mean squared error $N^{-1} \sum_i (p_i - y_i)^2$ [5]. Negative log-likelihood used the Bernoulli log loss. Expected calibration error divided $[0,1]$ into ten equal-width bins and computed the sample-weighted absolute gap between mean confidence and empirical accuracy [7], [13]. Adaptive ECE repeated the calculation with ten equal-mass bins, and maximum calibration error reported the largest equal-width gap. The mean confidence was also compared directly with forced-answer accuracy to identify global underconfidence or overconfidence.

$$\text{Brier} = (1/N) \sum_i (p_i - y_i)^2$$

$$\text{ECE} = \sum_\beta (n_\beta/N) |\text{accuracy}(B_\beta) - \text{confidence}(B_\beta)|$$

Selective metrics. Questions were sorted from highest to lowest confidence. At coverage c , the first $\text{ceiling}(cN)$ questions were accepted, and selective accuracy was the fraction correct among them. Selective risk equals one minus selective accuracy. The area under the risk-coverage curve, AURC, averaged cumulative risk across all attainable coverages; E-AURC subtracted the risk obtained by an oracle ordering that places every correct answer before every error. This formulation follows the risk-coverage view of selective prediction [21], [23], [24]. Correctness AUROC measured the ranking of correct versus incorrect selections independently of a particular coverage.

Cost-sensitive abstention. For an accepted wrong answer, the decision cost was 1. For an abstention, the cost λ took values 0.05, 0.10, 0.20, 0.30, 0.40, and 0.50. Each fold selected the confidence threshold that minimized $(\text{wrong accepted} + \lambda \times \text{times abstained})/N$ on its calibration questions. The threshold was then evaluated on the outer test fold. This rule is the finite-sample counterpart of reject-option decision theory [25]. Reporting several λ values avoids treating one operational cost as universal.

Statistical analysis. Exact paired McNemar tests compared selection errors. Accuracy, Brier, and AURC used 2,000 paired question-level bootstrap resamples [57], keeping each question's candidates and confidence variants together. Category intervals were omitted where counts were too small. This separation of logged data, selection, and held-out evaluation parallels counterfactual ranking, where naïve replay is policy-biased [58].

Implementation and artifacts. The experiment used Python, NumPy, pandas, SciPy, scikit-learn, and Matplotlib. Every table in the results section is derived from saved question-level or candidate-level predictions. The fixed seed controls answer shuffling, fold assignment, model fitting, and bootstrap sampling. The complete execution writes the transformed candidate table, fold predictions, reliability bins, metric tables, figures, software versions, and a cryptographic hash of the TruthfulQA data file.

Results and Discussion

Table Iii
Candidate truthfulness scorer performance on held-out questions.

Scorer	Accuracy	Mean confidence	Brier	ECE	AURC	Correctness AUROC
Fold-tuned ensemble	72.22%	0.540	0.201	0.183	0.118	0.781
Hybrid	71.11%	0.637	0.171	0.085	0.123	0.788
QA word + character	70.50%	0.489	0.217	0.217	0.124	0.786
Answer word + character	70.13%	0.558	0.194	0.145	0.132	0.776
Answer word	69.77%	0.508	0.215	0.190	0.144	0.765

Note: Accuracy is MC1-style best-answer selection. Confidence metrics use each scorer's raw within-question softmax.

Candidate selection. The fold-tuned ensemble selected the best answer on 72.22% of the 817 held-out questions. It exceeded the answer-word scorer by 2.45 percentage points, the answer word-plus-character scorer by 2.08 points, the question-answer scorer by 1.71 points, and the hybrid scorer by 1.10 points. The gain over the question-answer scorer was supported by the exact paired test (22 ensemble-only correct cases versus 8 baseline-only correct cases, $p=0.016$). The comparison with the hybrid scorer did not reach significance (25 versus 16, $p=0.211$), which is consistent with the modest absolute difference.

The fixed hybrid scorer produced the lowest raw Brier Score among individual scorers, but its accuracy was lower than the fold-tuned ensemble. This divergence illustrates why candidate selection and calibration must be evaluated separately. The ensemble weight varied from 0.500 to 0.875 across folds, indicating that both semantic pair features and hybrid structural features contributed. A single global weight would have ignored that fold-level variation.

Table Iv
Outer-fold diagnostics for selection and calibration.

Fold	QA-score weight	Temperature	Test accuracy	Raw Brier	CASA Brier
1	0.750	0.459	72.56%	0.210	0.135
2	0.625	0.775	75.00%	0.202	0.128
3	0.875	0.609	67.48%	0.183	0.140
4	0.500	0.701	74.23%	0.207	0.152
5	0.750	0.425	71.78%	0.201	0.162

Fold stability. Test accuracy ranged from 67.48% to 75.00%. CASA-Truth reduced Brier Score in every fold, with fold-specific values between 0.1277 and 0.1619. The fitted temperatures ranged from 0.425 to 0.775, all below one. Because temperature below one sharpens score differences, this pattern confirms that the raw within-question softmax was globally underconfident rather than overconfident: mean raw confidence was 53.97% while observed accuracy was 72.22%.

Calibration quality. CASA-Truth achieved the best Brier Score (0.1433), negative log-likelihood (0.4319), ECE (0.0312), AURC (0.0954), and correctness AUROC (0.8379). Relative to raw softmax, the Brier Score fell by 28.6%, ECE fell by 82.9%, and AURC fell by 18.8%. Mean CASA-Truth confidence was 71.78%, only 0.44 percentage points below accuracy.

Temperature scaling was the strongest conventional baseline for Brier Score (0.1629) and AURC (0.1088). Platt, isotonic, and beta calibration substantially corrected the global confidence level, yielding ECE between 0.0368 and 0.0418, but their AURC values remained between 0.1206 and 0.1284. These mappings mainly recalibrated a one-dimensional confidence score; they did not add enough information to reorder difficult cases. CASA-Truth improved both probability accuracy and ranking because it used score dispersion, candidate-set structure, lexical evidence, and question type.

Table V
Calibration and selective-ranking metrics on all outer test questions.

Method	Accuracy	Mean confidence	Brier	NLL	ECE	Adaptive ECE	AURC	Correctness AUROC
CASA-Truth	72.22%	71.78%	0.143	0.432	0.031	0.030	0.095	0.838
Temperature scaling	72.22%	65.67%	0.163	0.486	0.072	0.073	0.109	0.803
Platt scaling	72.22%	73.63%	0.163	0.487	0.030	0.032	0.121	0.777
Beta calibration	72.22%	74.06%	0.164	0.505	0.041	0.040	0.122	0.774
Isotonic regression	72.22%	74.04%	0.166	0.592	0.042	0.032	0.128	0.766
Raw softmax	72.22%	53.97%	0.201	0.575	0.183	0.182	0.118	0.781

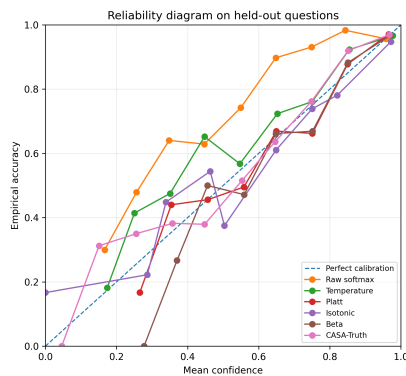


Fig. 4. Reliability diagram for raw and calibrated confidence methods.

The reliability diagram shows the same distinction visually. Raw confidence lies below the diagonal through most occupied bins, matching the underconfidence identified by temperatures below one. Temperature scaling moves the curve toward the diagonal but retains residual gaps. Platt and beta mappings align average confidence well but compress many cases into a narrower confidence range. CASA-Truth follows

the diagonal across the densely populated region and still separates low-confidence errors from high-confidence correct selections. Its maximum calibration error was 0.1616, compared with 0.2923 for raw softmax.

Calibration also needs effective presentation. Financial dashboards [59], progressive document rendering [60], visual evidence cards [61], and compact quality tokens [62] each prioritize decision-relevant information. CASA-Truth's reliability and coverage plots similarly expose how confidence, observed accuracy, and accepted workload interact.

Selective answering. At full coverage, all methods share the same 72.22% forced-answer accuracy because calibration does not alter the selected candidate. Differences emerge as coverage decreases. CASA-Truth reached 85.49% accuracy at 70.01% coverage and 93.40% accuracy at 50.06% coverage. At 30.11% coverage, accuracy was 97.97%; at 20.07% coverage, it was 98.17%. The raw score reached 81.82%, 89.24%, 95.94%, and 96.95% at the corresponding coverages.

Temperature scaling improved the accuracy-coverage curve relative to raw confidence because each fold used a different fitted temperature. Although temperature is monotonic within a fold, cross-fold temperature differences change the global ranking when all outer predictions are pooled. CASA-Truth remained stronger: at 60.10% coverage it answered 90.22% correctly, compared with 86.97% for temperature scaling and 85.13% for raw confidence. The risk-coverage figure expresses the same result as lower cumulative error.

Table Vi
Accuracy and risk at fixed coverage levels.

Method	Realized coverage	Selective accuracy	Selective risk
Raw softmax	100.00%	72.22%	27.78%
Raw softmax	80.05%	79.36%	20.64%
Raw softmax	70.01%	81.82%	18.18%
Raw softmax	60.10%	85.13%	14.87%
Raw softmax	50.06%	89.24%	10.76%
Raw softmax	30.11%	95.93%	4.07%
Raw softmax	20.07%	96.95%	3.05%
Temperature scaling	100.00%	72.22%	27.78%
Temperature scaling	80.05%	80.43%	19.57%
Temperature scaling	70.01%	82.69%	17.31%
Temperature scaling	60.10%	86.97%	13.03%
Temperature scaling	50.06%	91.20%	8.80%
Temperature scaling	30.11%	96.75%	3.25%
Temperature scaling	20.07%	97.56%	2.44%
CASA-Truth	100.00%	72.22%	27.78%
CASA-Truth	80.05%	81.50%	18.50%
CASA-Truth	70.01%	85.49%	14.51%
CASA-Truth	60.10%	90.22%	9.78%
CASA-Truth	50.06%	93.40%	6.60%
CASA-Truth	30.11%	97.97%	2.03%
CASA-Truth	20.07%	98.17%	1.83%

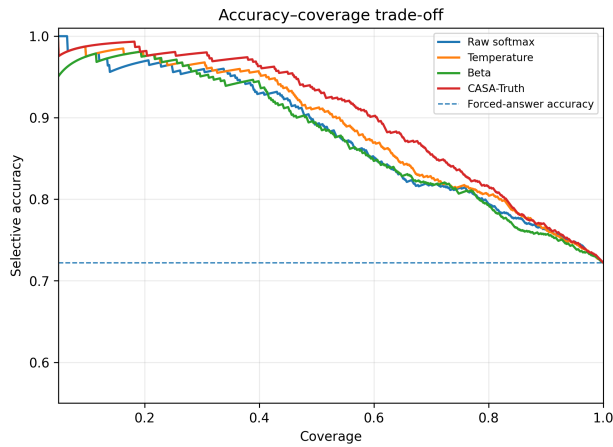


Fig. 5. Accuracy-coverage curves for forced and selective answering.

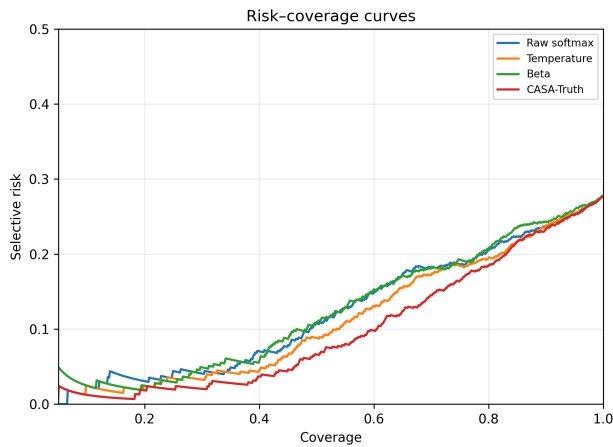


Fig. 6. Risk-coverage curves for confidence-based abstention.

Table VII

Maximum coverage while meeting target empirical accuracy.

Method	Target accuracy	Coverage	Achieved accuracy
Raw softmax	80.00%	78.95%	80.00%
Raw softmax	85.00%	60.10%	85.13%
Raw softmax	90.00%	48.96%	90.00%
Temperature scaling	80.00%	82.13%	80.03%
Temperature scaling	85.00%	64.01%	85.09%
Temperature scaling	90.00%	53.98%	90.02%
CASA-Truth	80.00%	82.99%	80.09%
CASA-Truth	85.00%	70.75%	85.12%
CASA-Truth	90.00%	60.47%	90.08%

Target-accuracy policies provide a deployment-oriented reading of these curves. CASA-Truth retained 82.99% coverage at an 80% target, 70.75% at an 85% target, and 60.47% at a 90% target. The corresponding raw-confidence coverages were 78.95%, 60.10%, and 48.96%. Thus the feature-aware confidence accepted 94 more questions than raw confidence while sustaining approximately 90% accuracy.

Confidence quartiles confirm that this gain is not produced by a narrow threshold artifact. The lowest CASA-Truth quartile had 38.05% accuracy and mean confidence 39.03%. Accuracy rose to 64.22% in the second quartile, 88.73% in the third, and

98.04% in the highest quartile. Mean confidence in the highest quartile was 98.00%. The monotonic progression shows that the calibrated score is simultaneously interpretable as a probability and useful as an error-ranking signal.

Table VIII

Observed outcomes across CASA-Truth confidence quartiles.

Confidence quartile	N	Mean confidence	Accuracy	Adversarial share	Mean candidates
Q1 (lowest)	205	39.03%	38.05%	40.00%	5.32
Q2	204	64.48%	64.22%	52.94%	5.28
Q3	204	85.77%	88.73%	64.22%	4.95
Q4 (highest)	204	98.00%	98.04%	56.86%	4.96

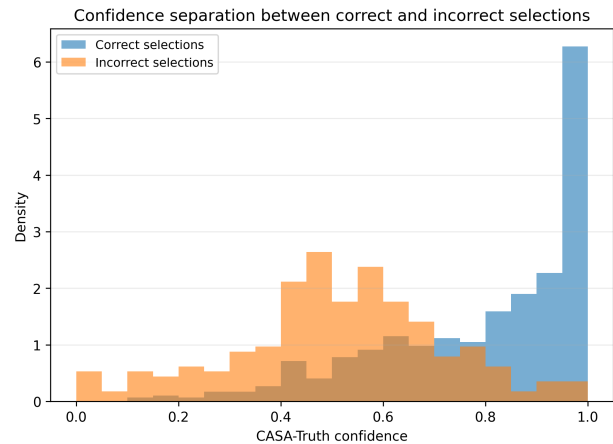


Fig. 7. Distribution of CASA-Truth confidence for correct and incorrect selections.

Subgroup behavior. Adversarial questions were answered more accurately than non-adversarial questions, 76.66% versus 67.11%. The adversarial set contains many explicit misconception and negation patterns, which are accessible to answer-form and lexical features. CASA-Truth reduced Brier Score in both groups, from 0.2074 to 0.1371 for adversarial questions and from 0.1930 to 0.1504 for non-adversarial questions. At 70% within-group coverage, selective accuracy reached 87.58% and 80.83%, respectively.

Category results expose where the scorer remains weak. Misconceptions, the largest category, achieved 54.00% forced accuracy. Conspiracies achieved 44.00%, History 58.33%, and Confusion: Places 20.00%. By contrast, Health reached 81.82%, Fiction 80.00%, Superstitions 95.45%, and Indexical Error: Other 95.24%. The confidence layer cannot recover a correct answer when candidate ranking fails, but it often identifies the failure: Confusion: Places had mean confidence 50.03% and Misconceptions 69.19%, both below the confidence of the strongest categories. Category ECE remains noisy at these sample sizes, so the category table is diagnostic rather than a separate benchmark leaderboard.

Table IX

Performance by TruthfulQA adversarial type.

Type	N	Accuracy	Raw Brier	CASA Brier	Raw ECE	CASA ECE	CASA AURC	Accuracy @70%
Adversarial	437	76.66%	0.207	0.137	0.223	0.028	0.082	87.58%
Non-Adversarial	380	67.11%	0.193	0.150	0.142	0.048	0.116	80.83%

Table X

Results for the twelve largest TruthfulQA categories.

Category	N	Forced accuracy	Mean confidence	Brier	AURC	Accuracy @70%
Misconceptions	100	54.00%	69.19%	0.208	0.229	68.57%
Law	64	76.56%	73.79%	0.140	0.075	86.67%
Health	55	81.82%	70.50%	0.096	0.026	100.00%
Sociology	55	67.27%	66.94%	0.136	0.103	84.62%
Economics	31	67.74%	62.81%	0.175	0.147	81.82%
Fiction	30	80.00%	79.36%	0.106	0.045	90.48%
Paranormal	26	84.62%	71.84%	0.150	0.066	84.21%
Conspiracies	25	44.00%	64.16%	0.198	0.280	55.56%
Stereotypes	24	75.00%	72.68%	0.163	0.097	88.24%
History	24	58.33%	51.88%	0.271	0.351	58.82%
Confusion: People	23	82.61%	66.28%	0.159	0.063	82.35%
Superstitions	22	95.45%	85.72%	0.071	0.007	100.00%

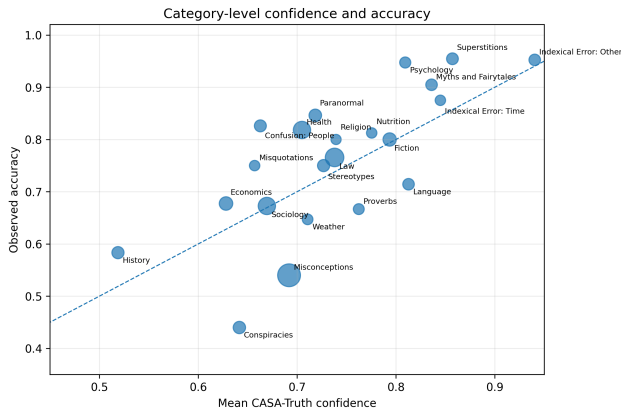


Fig. 8. Category-level observed accuracy and mean CASA-Truth confidence; marker area increases with category size.

Ablation analysis. Score-only feature-aware calibration reduced Brier Score from 0.2007 to 0.1556 and increased correctness AUROC from 0.7806 to 0.8015. Adding candidate-set size and length structure reduced Brier to 0.1534. Adding lexical overlap, negation, hedging, absolute terms, and explicit uncertainty reduced it further to 0.1449 and improved AURC from 0.1076 to 0.0968. The adversarial-type indicator produced the final Brier of 0.1433 and AURC of 0.0954. The largest improvement therefore came from using a learned multivariate calibration function; lexical features supplied the largest incremental gain after score features.

The ablation also clarifies why conventional one-dimensional mappings were insufficient. Raw top probability and margin summarize score concentration, but they do not distinguish a concise denial from an absolute misconception, a long answer from a short answer, or a two-candidate question from a thirteen-candidate question. CASA-Truth does not treat these signals as direct proof of truth. It uses their empirical

association with correctness within the calibration fold and regularizes the mapping before applying it to unseen questions.

Table Xi
CASA-Truth calibration feature ablations.

Ablation	Brier	NLL	ECE	Adaptive ECE	AURC	E-AURC	Correctness AUROC
Full CASA-Truth	0.1433	0.432	0.031	0.030	0.095	0.052	0.838
Score + structure + lexical	0.1450	0.436	0.030	0.034	0.097	0.054	0.834
Score + structure	0.1530	0.461	0.046	0.036	0.108	0.065	0.809
Score features	0.1560	0.466	0.038	0.039	0.110	0.067	0.802
Raw softmax	0.2010	0.575	0.183	0.182	0.118	0.075	0.781

Cost-sensitive policies. CASA-Truth achieved the lowest mean test cost at every evaluated abstention cost. At $\lambda=0.10$, its cost was 0.0758, compared with 0.0845 for raw softmax and 0.0842 for temperature scaling. At $\lambda=0.20$, CASA-Truth lowered cost by 10.29% relative to raw confidence while answering 55.82% of questions at 91.48% selective accuracy. At $\lambda=0.40$, it achieved cost 0.2186, coverage 69.51%, and selective accuracy 86.80%, whereas raw confidence achieved cost 0.2522 and selective accuracy 78.36%.

Low abstention cost favors a conservative policy. At $\lambda=0.05$, CASA-Truth answered 35.12% of questions with 97.69% selective accuracy, accepting seven wrong answers across the five test folds. As abstention became more expensive, coverage rose smoothly to 74.78% at $\lambda=0.50$. The cost curve therefore offers a transparent operating-point choice: a high-stakes system can use the left side of the curve, while a lower-stakes assistant can accept more questions.

Table Xii
Cost-sensitive policies selected on calibration folds.

Method	Abstention cost	Mean test cost	Mean coverage	Selective accuracy
CASA-Truth	0.10	0.076	42.59%	96.20%
CASA-Truth	0.20	0.139	55.82%	91.48%
CASA-Truth	0.30	0.190	63.01%	88.70%
CASA-Truth	0.40	0.219	69.51%	86.80%
CASA-Truth	0.50	0.249	74.78%	84.24%
Raw softmax	0.10	0.084	32.69%	95.49%
Raw softmax	0.20	0.154	41.13%	92.59%
Raw softmax	0.30	0.204	51.54%	88.97%
Raw softmax	0.40	0.252	62.27%	78.36%
Raw softmax	0.50	0.275	66.18%	76.62%
Temperature scaling	0.10	0.084	34.17%	95.44%
Temperature scaling	0.20	0.146	47.75%	91.93%
Temperature scaling	0.30	0.200	56.21%	88.74%
Temperature scaling	0.40	0.238	66.26%	81.59%
Temperature scaling	0.50	0.273	71.40%	78.76%

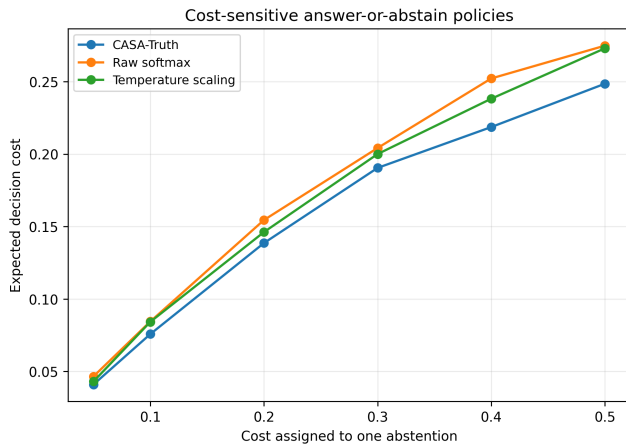


Fig. 9. Expected decision cost across abstention costs. Uncertainty of the reported improvements. The paired bootstrap 95% interval for forced-answer accuracy was [69.16%, 75.28%]. More importantly for the proposed contribution, the Brier reduction was 0.0574 with interval [0.0427, 0.0711], and the AURC reduction was 0.0222 with interval [0.0119, 0.0329]. Both paired intervals remained strictly above zero.

Taken together, the results support a specific conclusion rather than a general claim that confidence calibration solves hallucination. The candidate scorer still makes 227 errors at full coverage. Calibration changes what the system does with those errors: it identifies a large subset of difficult questions, assigns confidence values that match empirical frequencies, and permits thresholds that trade coverage for measurable risk. This distinction is operationally important because a truthful system must know both what to answer and when not to answer.

Table Xiii

Paired question-level bootstrap intervals (2,000 resamples).

Statistic	Estimate	95% CI lower	95% CI upper
Fold-tuned ensemble accuracy	0.7222	0.6916	0.7528
Raw Brier	0.2007	0.1875	0.2141
CASA-Truth Brier	0.1433	0.1308	0.1556
Brier reduction (raw minus CASA)	0.0574	0.0427	0.0711
Raw AURC	0.1176	0.0962	0.1410
CASA-Truth AURC	0.0954	0.0775	0.1156
AURC reduction (raw minus CASA)	0.0222	0.0119	0.0329

Limitations

The primary task evaluates a candidate-answer decision layer rather than unconstrained generation. TruthfulQA's best and incorrect answers define a controlled set in which selection accuracy and confidence are directly measurable. A free-form system must additionally generate candidate answers, normalize paraphrases, and determine whether a response is partially correct. The present results therefore characterize the answer-selection and abstention layer that follows generation, not the full end-to-end behavior of a specific generative backbone.

The candidate scorer is deliberately lightweight. TF-IDF, logistic regression, and answer-form features permit complete

five-fold evaluation on a compact benchmark and make every decision inspectable, but they do not contain the broad world knowledge of a modern language model. A stronger semantic scorer could increase full-coverage accuracy. The calibration findings remain relevant because the same fold-separated procedure can fit CASA-Truth to logits, self-evaluation scores, retrieval evidence, or consistency features from a larger backbone.

TruthfulQA contains 817 English questions and uneven category sizes. Calibration metrics for the full dataset are stable enough for paired evaluation, but category-level ECE values are sensitive to small counts. The category table therefore uses a minimum of fifteen questions and is interpreted as error analysis. Results on multilingual questions, rapidly changing facts, long-form answers, and domains with different misconception patterns require separate calibration data.

The calibration and test folds share one benchmark distribution. Shift can degrade uncertainty estimates [30]; approximately shared-feature methods offer one transfer strategy [63]. Production systems should still monitor drift and retain conservative fallbacks. The adversarial-type feature also assumes that label is available, but removing it changes Brier only from 0.1433 to 0.1449.

The abstention costs are normalized analytical values rather than costs from one deployment. Wrong medical advice, an incorrect historical fact, and a harmless fictional answer do not carry the same consequence. The cost curves expose the policy response across six values, but an application must set lambda from domain-specific harm, review capacity, latency, and user expectations.

C. CONCLUSION

CASA-Truth combines candidate truthfulness scoring, probability calibration, and selective abstention in one leakage-controlled evaluation. Across all 817 TruthfulQA questions, the fold-tuned scorer achieved 72.22% forced-answer accuracy. The feature-aware calibrator reduced Brier Score from 0.2007 to 0.1433, reduced ECE from 0.1826 to 0.0312, and reduced AURC from 0.1176 to 0.0954. These improvements appeared in every outer fold and remained positive under paired bootstrap sampling.

Selective answering converted those reliability gains into usable operating points. The system reached 85.49% accuracy at approximately 70% coverage and 93.40% accuracy at approximately 50% coverage. It maintained at least 90% empirical accuracy over 60.47% of questions and achieved the lowest cost across all tested abstention penalties. The central result is that truthfulness control benefits from treating confidence as a learned probability of correctness rather than a raw score.

For an LLM application, the practical architecture is direct: generate or retrieve candidate answers, score them, calibrate

the selected answer with held-out question-level data, and abstain when the expected error exceeds the cost of deferral. Accuracy, calibration, and coverage must be reported together because each describes a different part of trustworthy behavior.

D. REFERENCES

- [1] T. B. Brown et al., "Language models are few-shot learners," in *Advances in Neural Information Processing Systems* 33, 2020, pp. 1877-1901.
- [2] L. Ouyang et al., "Training language models to follow instructions with human feedback," in *Advances in Neural Information Processing Systems* 35, 2022, pp. 27730-27744.
- [3] 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.
- [4] S. Lin, J. Hilton, and O. Evans, "TruthfulQA: Measuring how models mimic human falsehoods," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, Dublin, Ireland, 2022, pp. 3214-3252.
- [5] G. W. Brier, "Verification of forecasts expressed in terms of probability," *Monthly Weather Review*, vol. 78, no. 1, pp. 1-3, 1950.
- [6] H. Murphy, "A new vector partition of the probability score," *Journal of Applied Meteorology*, vol. 12, no. 4, pp. 595-600, 1973.
- [7] Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. 34th Int. Conf. Machine Learning (ICML)*, vol. 70, 2017, pp. 1321-1330.
- [8] 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.
- [9] 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.
- [10] Niculescu-Mizil and R. Caruana, "Predicting good probabilities with supervised learning," in *Proc. 22nd Int. Conf. Machine Learning (ICML)*, 2005, pp. 625-632.
- [11] Zadrozny and C. Elkan, "Transforming classifier scores into accurate multiclass probability estimates," in *Proc. 8th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2002, pp. 694-699.
- [12] J. C. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *Advances in Large Margin Classifiers*, A. J. Smola, P. Bartlett, B. Schoelkopf, and D. Schuurmans, Eds. Cambridge, MA, USA: MIT Press, 1999, pp. 61-74.
- [13] M. P. Naeni, G. Cooper, and M. Hauskrecht, "Obtaining well calibrated probabilities using Bayesian binning," in *Proc. 29th AAAI Conf. Artificial Intelligence*, 2015, pp. 2901-2907.
- [14] M. Kull, T. Silva Filho, and P. Flach, "Beta calibration: A well-founded and easily implemented improvement on logistic calibration for binary classifiers," in *Proc. 20th Int. Conf. Artificial Intelligence and Statistics (AISTATS)*, vol. 54, 2017, pp. 623-631.
- [15] 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.
- [16] 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, Jan. 2024, doi: 10.54097/zJ4MnbWW.
- [17] 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.
- [18] Z. Ling, Q. Xin, Y. Lin, G. Su, and Z. Shui, "Optimization of autonomous driving image detection based on RFACONV and triplet attention," in *Proc. 2nd Int. Conf. Softw. Eng. Mach. Learn. (SEML)*, 2024, pp. 210-217, doi: 10.54254/2755-2721/77/2024MA0067.
- [19] 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.
- [20] 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.
- [21] Y. Geifman and R. El-Yaniv, "Selective classification for deep neural networks," in *Advances in Neural Information Processing Systems* 30, 2017, pp. 4885-4894.
- [22] 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.
- [23] Y. Geifman and R. El-Yaniv, "SelectiveNet: A deep neural network with an integrated reject option," in *Proc. 36th Int. Conf. Machine Learning (ICML)*, vol. 97, 2019, pp. 2151-2159.
- [24] R. El-Yaniv and Y. Wiener, "On the foundations of noise-free selective classification," *Journal of Machine Learning Research*, vol. 11, no. 53, pp. 1605-1641, 2010.
- [25] C. K. Chow, "On optimum recognition error and reject tradeoff," *IEEE Transactions on Information Theory*, vol. 16, no. 1, pp. 41-46, 1970.

- [26] 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.
- [27] Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," in *Proc. 5th Int. Conf. Learning Representations (ICLR)*, 2017.
- [28] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *Proc. 33rd Int. Conf. Machine Learning (ICML)*, vol. 48, 2016, pp. 1050-1059.
- [29] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Advances in Neural Information Processing Systems 30*, 2017, pp. 6402-6413.
- [30] Y. Ovadia et al., "Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift," in *Advances in Neural Information Processing Systems 32*, 2019, pp. 13991-14002.
- [31] 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.
- [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] 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.
- [34] Z. S. Zhong and S. Ling, "Uncertainty quantification of spectral estimator and MLE for orthogonal group synchronization," *arXiv:2408.05944*, Aug. 2024, doi: 10.48550/arXiv.2408.05944.
- [35] S. Desai and G. Durrett, "Calibration of pre-trained transformers," in *Proc. 2020 Conf. Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 295-302.
- [36] Z. Jiang, J. Araki, H. Ding, and G. Neubig, "How can we know when language models know? On the calibration of language models for question answering," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 962-977, 2021.
- [37] S. Lin, J. Hilton, and O. Evans, "Teaching models to express their uncertainty in words," *Transactions on Machine Learning Research*, 2022.
- [38] S. Kadavath et al., "Language models (mostly) know what they know," *arXiv:2207.05221*, 2022.
- [39] L. Kuhn, Y. Gal, and S. Farquhar, "Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2023.
- [40] 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 (EMNLP)*, Singapore, 2023, pp. 9004-9017.
- [41] 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.
- [42] 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.
- [43] 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.
- [44] 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.
- [45] 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, Apr. 2025, doi: 10.51903/jtie.v4i1.484.
- [46] K. 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. Comput. Data Sci. (ICCDs)*, 2024, pp. 143-149, doi: 10.54254/2755-2721/64/20241365.
- [47] 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.
- [48] 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.
- [49] 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.
- [50] 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.

- [51] B. 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. Comput. Data Sci. (CDS), 2024, pp. 94-99, doi: 10.54254/2755-2721/64/20241350.
- [52] 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. Comput. Data Sci. (CDS), 2024, pp. 64-70, doi: 10.54254/2755-2721/69/20241511.
- [53] G. S. Raj, H. Zhang, V. Abhyankar, S. Mukerjee, E. Zhang, J. Williams, R. Halverson, and J. M. Patel, "Impact of bilingual CS education on student learning and engagement in a data structures course," in Proc. 19th Koli Calling Int. Conf. Comput. Educ. Res., Koli, Finland, 2019, pp. 1-10, doi: 10.1145/3364510.3364518.
- [54] 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. iaie020, Sep. 2024, doi: 10.1093/imaia/iaie020.
- [55] Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," Information Processing & Management, vol. 24, no. 5, pp. 513-523, 1988.
- [56] Pedregosa et al., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825-2830, 2011.
- [57] B. Efron and R. J. Tibshirani, An Introduction to the Bootstrap. New York, NY, USA: Chapman & Hall, 1993.
- [58] 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.
- [59] 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.
- [60] 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.
- [61] 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.
- [62] 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.
- [63] Z. S. Zhong, X. Pan, and Q. Lei, "Bridging domains with approximately shared features," in Proc. 28th Int. Conf. Artif. Intell. Statist. (AISTATS), PMLR, vol. 258, 2025, pp. 559-567.