Stout in Mathematics and Statistics (SMS)

DOI: https://doi.org/10.61424/ka6yps02
Journal Homepage: https://stoutjournals.org/index.php/SMS

Distribution-Free Selective Prediction for LLM-RAG Hallucination Detection: Word-Level Conformal Risk Control under Domain Shift

Monica Lin
Computer Science, Michigan State University, East Lansing, MI, USA
Corresponding Author: Monica Lin, **E-mail**: m.lin.cs@gmail.com

ARTICLE INFORMATION

Submitted: April 05, 2026
Accepted: June 29, 2026
Published: July 31, 2026
Volume: 2
Issue: 1

KEYWORDS

Retrieval-augmented generation; hallucination detection; word-level classification; conformal prediction; selective prediction; probability calibration; domain shift; RAGTruth.

ABSTRACT

Retrieval-augmented generation (RAG) gives language models access to external evidence, yet a generated answer can still contain unsupported or contradictory words. This study evaluates word-level hallucination detection and uncertainty control on the complete RAGTruth corpus: 17,790 responses derived from 2,965 evidence sources across question answering, summarization, and data-to-text generation. To prevent evidence leakage, source identifiers were partitioned before any response or token was assigned to fit, development, calibration, and test sets. Three computationally light detectors were trained and evaluated: a deterministic evidence-overlap rule, a 131,072-dimensional hashed logistic model, and a compact multilayer perceptron using 55 dense lexical, local-context, and evidence-alignment features. Platt calibration converted detector scores to probabilities, after which pooled split conformal, task-Mondrian conformal, and source-block maximum calibration produced binary prediction sets at nominal coverage levels of 0.95, 0.90, and 0.80. The hashed logistic detector achieved the strongest thresholded token F1 (0.294), compared with 0.270 for the compact MLP and 0.117 for the rule, while the MLP achieved the highest AUROC (0.813) and lowest expected calibration error (0.0099). At 90% nominal coverage, pooled and task-Mondrian prediction sets obtained 0.917 and 0.918 empirical coverage with mean set sizes of 0.945 and 0.949. Source-block maximum calibration covered every test token but always returned both labels, demonstrating the cost of conservative cluster protection. Leave-one-task-out experiments revealed substantial shift sensitivity: empirical coverage ranged from 0.752 on held-out data-to-text generation to 0.970 on held-out summarization. The results establish a reproducible, source-disjoint benchmark for selective word-level RAG auditing and clarify the boundary between finite-sample marginal coverage under exchangeability and empirical robustness under task shift.

1. Introduction

Retrieval-augmented generation couples parametric generation with an external evidence store so that a system can answer knowledge-intensive questions with information available at inference time (Lewis et al., 2020). Retrieval often reduces unsupported generation, but it does not remove it: a model can ignore the retrieved record, merge incompatible statements, introduce an unsupported qualifier, or state a conclusion that the evidence does not license (Shuster et al., 2021). For evidence-sensitive use, the practical question is therefore not only whether an answer is fluent, but which words can be trusted and when the detector should withhold a definitive label.

Evidence attribution is especially important in long answers. Citation-bearing text can look grounded even when a citation fails to support a nearby claim, and aggregate answer scores can conceal a small but consequential unsupported span (Gao et al., 2023). RAGTruth was designed for this gap. It contains manually annotated responses generated from question-answering passages, summarization documents, and structured data records, with both response-level and word-level hallucination labels (Niu et al., 2024). Its three tasks and six generators make it possible to study localization, probability calibration, and transfer rather than reducing hallucination to a single answer-level decision.

A useful detector must also express uncertainty in an operational form. Selective classification permits a model to abstain on uncertain cases and evaluates the resulting trade-off between retained coverage and conditional error (El-Yaniv & Wiener, 2010). In language tasks, naive confidence thresholds can remain overconfident after a domain change, which makes held-out-domain assessment essential (Kamath et al., 2020). Conformal prediction offers a complementary construction: calibration nonconformity scores are converted into prediction sets with finite-sample marginal coverage when calibration and test units are exchangeable (Angelopoulos & Bates, 2023). The set can contain the supported label, the hallucinated label, both labels, or, for some score constructions, neither label. Set size and singleton frequency then quantify the practical cost of uncertainty.

This study asks four connected questions. First, how well can transparent and inexpensive models localize hallucinated words when all RAGTruth responses are used and evidence-source leakage is prevented? Second, how accurately do their probabilities reflect observed token risk? Third, what coverage-efficiency trade-offs arise from pooled, task-conditional, and source-block conformal calibration? Fourth, how much of that behavior persists when an entire RAG task is held out? These questions are answered with one fixed source-disjoint protocol, saved token predictions, twelve empirical tables, and eight figures generated from the executed run.

The contribution is methodological as much as predictive. The experiment separates thresholded localization metrics from prediction-set coverage; it reports both exact and overlap-based span scores; and it treats leave-one-task-out coverage as an empirical shift result rather than an automatic extension of the exchangeable-data theorem. This distinction matters because distribution-free marginal coverage does not imply object-conditional validity, span-F1 control, or unchanged validity after an arbitrary task shift (Vovk, 2012; Barber et al., 2023).

Two evaluation choices are particularly consequential. First, the positive class is rare at word level even when a response contains a hallucination. A detector that marks every word as supported can consequently achieve high accuracy while offering no auditing value. The analysis therefore places token F1, AUPRC, balanced accuracy, and MCC beside accuracy, and it preserves precision and recall so that false-alarm and missed-span costs remain visible. Second, the experimental unit is not an isolated token. Six responses share each evidence source, and tokens within a response are strongly dependent. Partitioning by source avoids direct leakage, while source-block calibration tests what happens when the dependence structure is handled through a worst-token aggregate. These decisions connect the statistical estimand to the way a deployed RAG auditor would encounter evidence and responses.

The word-level setting also provides a concrete interface for human review. A response-level warning forces a reviewer to reread the entire answer and evidence bundle; a localized warning directs attention to the phrase most likely to require correction. Conversely, excessive fragmentation can overwhelm the reviewer even when token recall is high. This motivates reporting predicted-span counts and exact as well as relaxed boundary scores. The study thus evaluates not just whether a model ranks suspicious tokens, but whether that ranking can support a selective workflow with interpretable uncertainty.

2. Literature Review

2.1 Fine-grained RAG evaluation and factuality

RAG evaluation has moved from a single answer-quality score toward decomposed diagnostics. RAGAs measures dimensions such as context relevance and faithfulness without requiring a task-specific human reference (Es et al., 2024), while ARES uses lightweight learned judges and examines behavior when query or document domains change (Saad-Falcon et al., 2024). RAGChecker further separates retrieval failures from generation failures using fine-grained metrics that expose the mechanism of an error (Ru et al., 2024). These frameworks motivate the present focus on evidence-conditioned word decisions: an answer may be generally relevant while containing a locally unsupported phrase.

The decomposition also clarifies what the detector can and cannot establish. Retrieval relevance asks whether the supplied record is useful for the query; answer relevance asks whether the response addresses the query; evidence faithfulness asks whether the response is licensed by that record. A word-level hallucination label is closest to the third construct. It does not prove that the evidence itself is true, nor does it evaluate whether a better document should have been retrieved. The present models therefore condition on the evidence already associated with each response and estimate local support or conflict. This narrower target permits faithful measurement without conflating the upstream retriever, the generator, and the evaluator.

Hallucination benchmarks differ in granularity and evidence assumptions. HaluEval spans general queries, question answering, dialogue, and summarization, showing that hallucination recognition remains difficult even when external knowledge and reasoning are available (J. Li et al., 2023). TRUE demonstrates that factual-consistency evaluators should be compared at the instance level across tasks because system-level correlations can hide important errors (Honovich et al., 2022). HaDeS supplies token-level supervision for reference-free generated text, directly motivating localization rather than a sentence-only label (T. Liu

et al., 2022). FaithDial shows a related need in knowledge-grounded dialogue and demonstrates transfer of a trained hallucination critic (Dziri et al., 2022).

Other work decomposes or self-checks generated content. SelfCheckGPT detects sentence-level non-factuality from disagreement among sampled continuations without using an external database (Manakul et al., 2023). FActScore decomposes long-form generations into atomic claims and scores factual precision at that finer level (Min et al., 2023). These approaches address broad factuality, whereas RAGTruth binds each annotation to a supplied evidence source and therefore supports direct measurement of evidence conflict and unsupported additions. The distinction is crucial: factual truth, textual entailment, evidence overlap, and citation completeness are related but not identical properties.

2.2 Abstention and probability calibration

Abstention has a parallel history in machine reading. SQuAD 2.0 requires a model to identify questions that cannot be answered from the passage, thereby treating non-answering as part of reliable prediction (Rajpurkar et al., 2018). SelectiveNet integrates a reject option into neural training at a target coverage, whereas post-hoc selection uses a separate confidence score after fitting (Geifman & El-Yaniv, 2019). Post-hoc selection is appropriate here because the same fitted detector can be evaluated under multiple coverage targets and calibration schemes.

Probability calibration is a prerequisite for interpretable selection but not a guarantee of transfer. Neural classifiers can be sharply miscalibrated, and simple post-hoc scaling is often effective in-domain (Guo et al., 2017). Pretrained language representations also show different calibration behavior in and out of domain (Desai & Durrett, 2020), while generative question-answering confidence often needs explicit calibration before it supports reliable abstention (Jiang et al., 2021). The present pipeline applies Platt scaling on the source-disjoint development partition and reports Brier score, negative log likelihood, and expected calibration error alongside discrimination metrics. The distinct calibration partition is then reserved for conformal quantiles, preventing the same labels from selecting both the probability map and the final prediction-set threshold.

2.3 Conformal prediction and shift

Conformal prediction converts calibration residuals or nonconformity scores into set-valued predictions. Adaptive classification sets seek small sets for easy cases while retaining marginal coverage (Romano et al., 2020). Risk-controlling prediction sets generalize the target from miscoverage to other expected losses (Bates et al., 2021), and conformal risk control extends this idea to bounded monotone losses over nested families, including a token-based natural-language setting (Angelopoulos et al., 2024). Conformal language modeling likewise shows how calibrated sets can represent uncertainty for open-ended outputs (Quach et al., 2024). In this experiment, the formal prediction object is the binary token-label set. Thresholded token F1 and span F1 are empirical localization metrics; they are not treated as if an arbitrary span-F1 threshold inherited the prediction-set coverage theorem.

Group validity and distribution shift require additional care. Mondrian calibration can provide validity within prespecified categories when exchangeability holds inside each category, although limited calibration counts can make sets unstable or broad (Vovk, 2012). Weighted conformal prediction can address covariate shift when reliable likelihood ratios are available (Tibshirani et al., 2019), and adaptive methods can target long-run coverage in changing sequences (Gibbs & Candès, 2021). Recent non-exchangeable risk-control procedures introduce weighted corrections for dependent or drifting data (Farinhas et al., 2024). Those corrections were not part of the executed pipeline; consequently, the leave-one-task-out results quantify observed robustness and do not claim exact shifted-domain validity. This boundary follows current surveys of conformal NLP, which emphasize structured outputs and shift as open efficiency and validity challenges (Campos et al., 2024).

2.4 Evidence chains, provenance, and retrieval reliability

Recent RAG research increasingly treats grounding as an auditable chain rather than a single relevance score. Word-level evidence-chain systems connect a questionable span to source attribution and provenance, while related work combines deterministic provenance with calibrated abstention for unanswerable questions (C. Li et al., 2024; Jin, 2025a; Zhong, Chen, et al., 2025). Operational studies reach the same conclusion from another direction: private-document DevOps retrieval, bilingual incident triage, budgeted multi-hop retrieval, retrieval-quality prediction, and multilingual humanitarian RAG all separate evidence access from the reliability of the generated answer (B. Zhang et al., 2024; G. Liu et al., 2024; Su et al., 2025; R. Zhang et al., 2025; Y. Chen & Xu, 2026). These studies support the present focus on local evidence support: a useful warning should identify the questionable words and preserve a traceable path to the record against which they were judged.

Domain-specific RAG further shows why retrieval quality, answerability, and evidence-constrained reasoning must remain distinct. Accounting due-diligence retrieval, hybrid financial search, numerical guardrails for quantitative assistants, and evidence-constrained monitoring of tokenized assets require generated statements to remain tied to disclosures or numerical

records (Y. Chen et al., 2025; Meng et al., 2026; Z. Li et al., 2026; S. Zhou et al., 2026). Legal-governance RAG and scientific claim verification add policy and narrative constraints, while a therapist-facing copilot makes source ranking visible to a professional user (J. Li & Zhou, 2026; Su, Chen, et al., 2026; Y. Zhang & H. Zhang, 2025a). None of these application studies establishes conformal validity for RAGTruth; together, they show that evidence support is a separate prediction target whose errors should be localized and audited.

Grounding also depends on how evidence is retained and selected over time. Confidence-gated memory writing and time-aware retrieval address stale or conflicting information in multi-session dialogue, while federated preference learning studies knowledge-grounded personalization under privacy constraints (Sun et al., 2023; Kuo et al., 2025). Review-grounded recommendation evaluates whether an explanation remains faithful to retrieved user evidence, and retrieval-summary integration in code intelligence distinguishes findability from explanatory usefulness (Chang et al., 2026; Y. Li, 2024). Cross-modal medical pretraining provides a complementary transfer example: a representation that works in one task still requires direct evaluation after its evidence form and label structure change (Mi et al., 2025). Taken together, the memory and recommendation studies motivate preserving evidence-source boundaries, while the cross-modal result reinforces direct evaluation after the operating distribution changes.

Operational reliability studies make the provenance requirement especially concrete. Multi-source root-cause attribution, retrieved normal prototypes for OpenStack, retrieval-grounded HDFS narratives, and conformal alert control all connect a score to an inspectable incident trace (G. Liu et al., 2023; Xin, 2025a; Sun et al., 2026; Xin, 2026e). Cost-aware LogEval routing, retrieval-augmented runbook generation, and incident-visualization cards similarly separate detection, evidence selection, explanation, and action (C. Li et al., 2026; B. Zhang et al., 2026; Nie et al., 2026). Trajectory-reliability prediction extends this logic from one answer to the sequence of tool-use decisions that produced it, while hybrid-cloud deployment makes model routing an explicit operational choice (Zhong et al., 2026; Xin, 2025b). The present detector is deliberately narrower, but its word probabilities and prediction sets can support the same evidence-first review pattern.

2.5 Calibration, rare events, and domain transfer

Calibrated action is not unique to hallucination detection. Credit-risk tiering uses uncertainty-driven rejection, probability-of-default workflows combine calibration with distribution-free uncertainty, and perception fusion learns how confidence should change when evidence streams disagree (Y. Chen et al., 2023; Jin et al., 2024b; Xin, 2025c). Medical vision-language classification and resume-job matching likewise distinguish ranking quality from the reliability of a probability used for selective review (Lu & Zou, 2026; B. Zhou, Jin, et al., 2025). In infrastructure, conformal demand envelopes translate predictive uncertainty into a bounded oversubscription decision (S. Chen et al., 2024). These cross-domain examples do not transfer their guarantees to this corpus, but they reinforce why AUPRC, calibration loss, selective error, conformal coverage, and set size answer different operational questions.

Rare-event detection poses a related measurement problem. Extreme-imbalance fraud and bankruptcy studies use cost sensitivity or resampling because raw accuracy can conceal failure on the positive class (Jin et al., 2024a; Y. Chen et al., 2024). Network and host-security studies similarly evaluate feature selection, compact IoT detection, deep-autoencoder traffic classification, DDoS mitigation, CloudTrail attack stages, system-call localization, and self-supervised log anomalies under sparse or structured positives (Y. Li & Lu, 2025; Lu & Zhou, 2024; Xin et al., 2024; Wang et al., 2024; Bai et al., 2026; Xin, 2026d; Xin, 2026g). Dark-pattern detection adds a text-classification example in which model outputs are paired with explanations (Xu et al., 2025). Together, these studies reinforce the need for metrics beyond raw accuracy; RAGTruth's word-level labels additionally make token F1 and span localization available for the present 4.20%-positive test set.

Distribution shift changes both probability meaning and conformal efficiency. Safe capacity forecasting, cross-cloud transfer, and few-shot cold-start forecasting confront heterogeneity in workloads, topology, or tenant history (He et al., 2023; He et al., 2024; He et al., 2025). Noisy-neighbor degradation and multi-horizon demand forecasting illustrate heterogeneity across hosts and workloads even when the prediction objective appears stable (Nie & Zheng, 2024; Zhao et al., 2025). Regime-aware bike-demand forecasting provides a temporal analogue, while shared-feature theory formalizes why only part of a representation may transfer across domains (Xin, 2026f; Zhong, Pan, et al., 2025). Spectral rank aggregation and uncertainty quantification for synchronization further separate estimator accuracy from an uncertainty statement about that estimator (Zhong & Ling, 2024a, 2024b). The leave-one-task-out experiment therefore measures transfer directly; it does not extend an exchangeability guarantee by analogy.

2.6 Human oversight, safety, and evidence interfaces

Selective prediction is most useful when its output can be interpreted and acted on. Scientific-search evidence cards, medical safety-card benchmarks, financial-risk dashboards, and capacity-dashboard briefs organize evidence, confidence, and escalation

cues instead of displaying an unexplained scalar score (Jin, 2025b; C. Li et al., 2025; Z. Li et al., 2025; G. Liu et al., 2025). Medical-image explanation cards and patient-facing risk cards likewise emphasize visible uncertainty and the separation of source evidence from generated explanation (Zhong, Wu, et al., 2025; B. Zhou, Li, et al., 2025). Recommendation cards, accounting-review cards, and an LLM design critic treat explanations as inspectable guidance rather than ground truth, while counseling-support cards keep evidence linked to professional review (B. Zhang et al., 2025; K. Zhang et al., 2025; Y. Li et al., 2025; Y. Zhang & H. Zhang, 2025b). For a binary word label, singleton, two-label, and empty conformal sets can support the same interface distinction between an automated decision and a case that requires inspection.

Safety work shows why abstention must be evaluated together with utility. Continual red-teaming updates guardrails as jailbreak behavior evolves, behavior-level refusal aims to preserve legitimate use, and VerifySafe inserts evidence-based self-verification before release (Zheng et al., 2023; Zheng & Li, 2024; Zheng et al., 2024). Privacy and data-integrity risk cards then make prompt-injection and oversight risks visible to a human operator, while strategy-aware emotional-support response control distinguishes explicit response strategies from unrestricted imitation (Su, Rao, et al., 2026; Y. Zhang & Z. Zhou, 2026). These studies concern adversarial or high-stakes settings rather than ordinary RAGTruth errors, but they support a conservative interpretation of non-singleton predictions: uncertainty should be surfaced rather than silently converted into a definitive label.

The same principle appears in educational and financial decision support. Rubric-grounded essay scoring ties feedback to explicit criteria, early-warning analytics couples a risk estimate with an intervention rationale, and counterfactual credit explanations expose what would have to change for a different decision (Xin, 2026a; Xin, 2026b; Xin, 2026c). Teacher-facing explanations similarly route predictions into a human decision process rather than treating model output as self-validating (J. Zhang, 2026). Accordingly, the present results support calibrated triage, not automatic deletion or rewriting of flagged text.

3. Methodology

3.1 Dataset and source-disjoint design

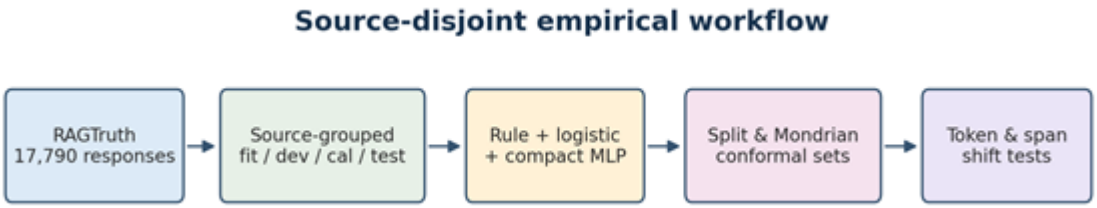
The experiment used the complete official RAGTruth response and source files distributed by Particle Media (n.d.). Each source instance belongs to question answering (QA), summarization, or data-to-text generation (Data2txt), and each source is associated with six generated responses. All 17,790 responses were included. Character-offset annotations were aligned to the deterministic word-token representation used by the analysis code; a token received the positive label when it intersected an annotated hallucination span. The pipeline retained annotation categories for sensitivity analysis while fitting one binary detector: supported versus hallucinated.

The official RAGTruth test split was retained, and source identifiers in the official training split were randomly assigned to fit, development, and calibration partitions. All six responses linked to a source remained together, and no source crossed a partition boundary. This design prevents nearly identical evidence content from appearing on both sides of evaluation and makes the source the natural clustering unit. The resulting test set contained 450 sources, 2,700 responses, and 342,424 word tokens. The fit partition contained 1,760 sources and 1,393,202 tokens; development contained 251 sources and 199,029 tokens; calibration contained 504 sources and 398,029 tokens.

The four partitions served non-overlapping roles. Fit tokens determined the trainable coefficients or neural weights. Development tokens selected the word-decision thresholds, fitted the Platt mappings, and stopped MLP training. Calibration tokens supplied only the nonconformity distributions used to select pooled, task-specific, or source-block quantiles. Test tokens remained untouched until the final calculation of discrimination, calibration, localization, selection, and coverage metrics. This separation prevents threshold optimization from using the same outcomes later invoked for conformal assessment.

The split retained the naturally occurring task and generator mixture rather than resampling to equal prevalence. That choice preserves the benchmark's observed operating conditions, but it also makes macro and subgroup views necessary. QA, Summary, and Data2txt have different evidence formats, response lengths, and positive rates; the six generators likewise produce markedly different token prevalence. Overall metrics therefore answer a population-weighted benchmark question, while task- and generator-specific metrics reveal heterogeneity that the aggregate can hide.

Figure 1. Executed source-disjoint pipeline from RAGTruth records to calibrated prediction sets, shift tests, tables, and figures.



3.2 Word-level detectors

The deterministic rule used eight evidence-overlap and surface indicators. It assigned higher suspicion to response tokens that lacked normalized lexical support in the supplied evidence, with stopword, token-form, and evidence-bigram adjustments. The rule has no trainable coefficients and establishes how far direct matching can go without statistical fitting.

The linear detector streamed token-centered lexical and local-context features through a 131,072-dimensional hashing map. It fitted 131,073 coefficients, including the intercept, in two passes over the fit tokens. Feature hashing bounded memory while preserving a large vocabulary of response, neighborhood, and evidence-alignment patterns. The development partition selected the decision threshold that maximized token F1; the selected threshold was 0.1582 after probability calibration.

The compact MLP consumed 55 dense features covering token form, local context, response position, normalized evidence overlap, and evidence-response alignment. Its two hidden layers contained 32 and 16 units, for 2,337 trainable parameters. Training stopped after 11 fitted epochs. This model tests whether nonlinear interactions among compact, inspectable features can match the much wider hashed representation. Its development-selected probability threshold was 0.1534.

3.3 Calibration and evaluation metrics

Raw detector scores were mapped to probabilities by a one-dimensional Platt model fitted without using test labels. Calibration quality was measured by Brier score, negative log likelihood, and expected calibration error. Discrimination was measured by area under the precision-recall curve (AUPRC) and area under the receiver-operating-characteristic curve (AUROC). Thresholded token metrics comprised precision, recall, F1, accuracy, balanced accuracy, and Matthews correlation coefficient. Because hallucinated words are rare, F1, AUPRC, balanced accuracy, and MCC are more informative than accuracy alone.

Each metric answers a different question. AUPRC evaluates ranking in the positive, imbalanced class; AUROC averages discrimination across positive and negative ranks; Brier score measures squared probability error; and negative log likelihood penalizes confident mistakes strongly. Expected calibration error groups predictions by confidence and aggregates the absolute difference between mean confidence and observed frequency. Token F1 is the harmonic mean of thresholded precision and recall. Balanced accuracy averages sensitivity and specificity, whereas MCC summarizes all four confusion-matrix cells and remains informative when class sizes differ. No single metric was used as a substitute for this joint view.

Consecutive positive word predictions were merged into spans. Exact span precision and recall required a predicted span to match a gold span boundary exactly. Relaxed span precision and recall used one-to-one matching and required token-span intersection over union of at least 0.50. The two views separate boundary accuracy from approximate localization without allowing one broad predicted span to claim several gold spans. A detector can achieve useful token ranking while producing fragmented spans, so token F1 and span F1 were never treated as interchangeable.

3.4 Conformal sets and selective prediction

For a calibrated hallucination probability p(x), the class probabilities were p1(x)=p(x) and p0(x)=1-p(x). For a calibration token with observed class y, nonconformity was s(x,y)=1-py(x). For target miscoverage alpha, the finite-sample upper empirical quantile of calibration scores produced q-alpha. The prediction set was C-alpha(x)={y in {0,1}: 1-py(x) <= q-alpha}. Coverage is the fraction of test tokens whose true label belongs to C-alpha(x); mean set size and singleton rate measure efficiency. Selective error is classification error on singleton sets only, while empty-rate and both-label rate describe the two non-singleton outcomes. This construction follows the practical split-conformal framework (Angelopoulos & Bates, 2023).

The finite-sample quantile uses the conformal rank correction rather than an unadjusted population percentile. Under exchangeability of calibration and test units, this rank construction yields marginal coverage for the true binary label. Marginal coverage averages across tokens; it does not promise the same coverage for every evidence source, response, task, generator, or individual token. Mean set size is consequently reported beside coverage. A method that always returns both binary labels has

perfect coverage but conveys no class decision, while a high singleton rate indicates that the calibrated set usually resolves to one label.

Pooled split conformal used one quantile across all calibration tokens. Task-Mondrian conformal calculated separate quantiles for QA, Summary, and Data2txt and applied each before aggregating test outcomes. Source-block maximum calibration first took a worst-case score within each evidence source and calibrated across those source maxima. It was designed to respect source clustering but is intentionally stringent: one difficult token can determine an entire source score. The three procedures were evaluated at alpha values 0.05, 0.10, and 0.20.

Risk-coverage curves swept decision confidence and measured error among retained token decisions. A separate leave-one-task-out (LOTO) protocol trained and calibrated the compact pipeline on two tasks and evaluated it on the third. The held-out task was absent from fitting, threshold selection, and conformal quantile estimation. The resulting coverage is empirical under task shift; exact split-conformal validity would require an exchangeability or shift-correction condition not established by the LOTO design. Generator-level results, annotation-type recall, feature ablations, and measured training and test time completed the robustness analysis.

The selective and conformal evaluations were kept conceptually separate. Selective coverage is the fraction of token decisions retained after uncertain cases are rejected, and selective error is computed only over those retained decisions. Prediction-set coverage instead asks whether the true label belongs to a set and can remain high even when the set contains both labels. This distinction prevents a low selective error from being mistaken for conformal validity and prevents a fully ambiguous two-label set from being described as an accurate point prediction.

3.5 Computational reproducibility

The run wrote source partitions, fitted models, thresholds, calibrator parameters, compressed word predictions, all table CSV files, and every plotted figure to disk. The manuscript values were read directly from those artifacts. The code starts from the two RAGTruth JSONL files, recreates the source-disjoint partitions, fits the three detectors, evaluates each test token and span, and regenerates the reported outputs.

4. Results and Discussion

4.1 Corpus and partitions

The 2,965 evidence sources produced exactly 17,790 responses. QA and Summary had similar response-level hallucination rates, 29.1% and 29.8%, whereas 68.6% of Data2txt responses contained at least one annotated hallucination. Across the corpus, 7,664 responses were positive and 14,289 spans were annotated. Evident baseless information was the largest annotation category (6,237 spans), followed by evident conflict (5,324) and subtle baseless information (2,527). Subtle conflict was uncommon (201 spans). This distribution explains why overall accuracy can remain high even when positive-token recall is modest.

Table 1. RAGTruth composition by task.

Task	Sources	Responses	Positive responses	Gold spans	Positive rate
QA	989	5,934	1,724	2,927	29.1%
Summary	943	5,658	1,686	2,072	29.8%
Data2txt	1,033	6,198	4,254	9,290	68.6%
Overall	2,965	17,790	7,664	14,289	43.1%

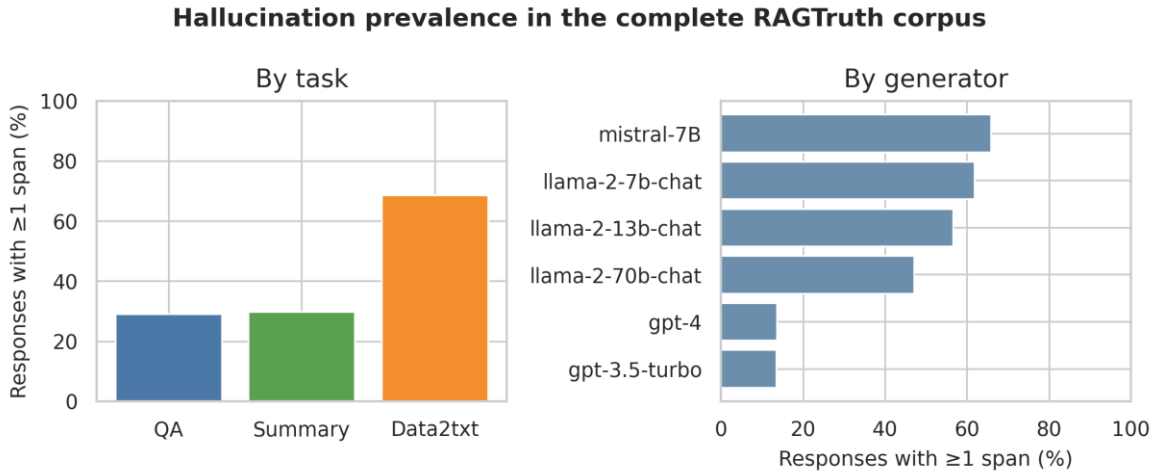
Table 2. Annotation-label distribution; the last two rows are overlapping analytical flags.

Category	Count	Share of spans
Evident Baseless Info	6,237	43.6%
Evident Conflict	5,324	37.3%
Subtle Baseless Info	2,527	17.7%
Subtle Conflict	201	1.4%
Implicitly true but unsupported spans	1,928	13.5%
Null-derived spans	1,642	11.5%

Table 3. Source-disjoint experimental partitions and token prevalence.

Partition	Responses	Sources	Word tokens	Positive tokens	Positive rate
fit	10,560	1,760	1,393,202	78,093	5.6%
dev	1,506	251	199,029	11,541	5.8%
calibration	3,024	504	398,029	22,625	5.7%
test	2,700	450	342,424	14,382	4.2%

Figure 2. Response-level hallucination prevalence by task and generator, computed from the complete corpus.



4.2 Detector discrimination and localization

The hashed logistic detector produced the highest thresholded token F1, 0.294, with precision 0.241 and recall 0.379. The compact MLP ranked tokens slightly better by AUROC (0.813 versus 0.798) but had lower F1 (0.270), reflecting a different operating-point trade-off. The deterministic rule reached recall 0.461 but only 0.067 precision, yielding F1 0.117 and 58,828 predicted spans. The logistic model reduced that count to 9,933 and achieved the strongest MCC (0.264). These results show that unsupported vocabulary overlap is useful as a screening signal but is too permissive for localization without learned context.

Span results were much lower than token results. For the logistic detector, exact span F1 was 0.023 and relaxed span F1 was 0.067; the corresponding MLP values were 0.015 and 0.052. The gap reflects boundary fragmentation: a few extra or missing words prevent exact agreement, and overlap scoring remains sensitive to the large number of short predicted fragments. Data2txt was the clearest domain for the logistic detector, with token F1 0.337 and relaxed span F1 0.151. Summary was hardest at the selected global threshold, with token F1 0.145 and recall 0.114. A single operating threshold therefore does not equalize behavior across tasks.

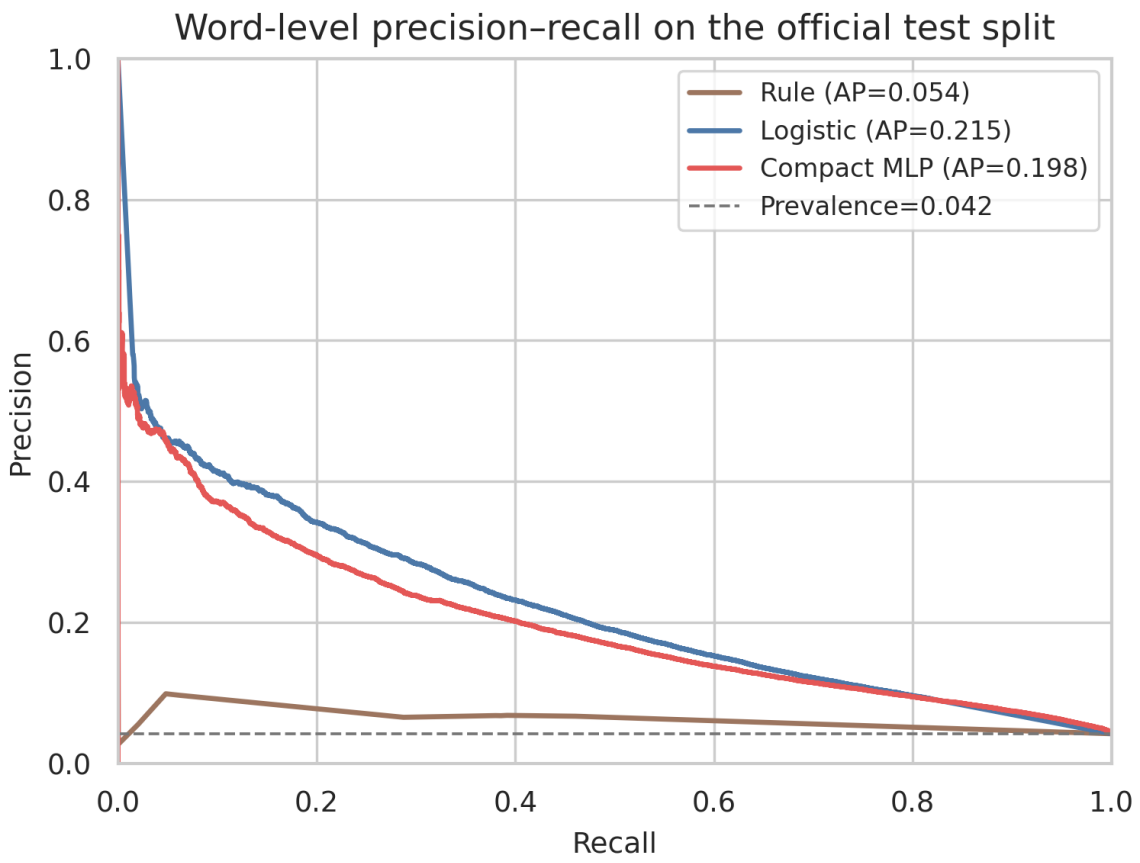
Table 4. Overall word-level and span-level test performance at development-selected thresholds.

Model	Precision	Recall	Token F1	AUPRC	AUROC	Relaxed span F1
Logistic	0.241	0.379	0.294	0.215	0.798	0.067
Compact MLP	0.212	0.369	0.270	0.198	0.813	0.052
Rule	0.067	0.461	0.117	0.054	0.591	0.012

Table 5. Task-specific test performance for the three detectors.

Task	Model	Precision	Recall	Token F1	AUROC	Relaxed span F1
Overall	Rule	0.067	0.461	0.117	0.591	0.012
Overall	Logistic	0.241	0.379	0.294	0.798	0.067
Overall	Compact MLP	0.212	0.369	0.270	0.813	0.052
QA	Rule	0.108	0.439	0.173	0.622	0.003
QA	Logistic	0.210	0.528	0.300	0.824	0.012
QA	Compact MLP	0.197	0.596	0.296	0.841	0.017
Summary	Rule	0.054	0.315	0.093	0.574	0.005
Summary	Logistic	0.200	0.114	0.145	0.698	0.018
Summary	Compact MLP	0.221	0.103	0.140	0.777	0.025
Data2txt	Rule	0.056	0.550	0.101	0.566	0.018
Data2txt	Logistic	0.304	0.378	0.337	0.817	0.151
Data2txt	Compact MLP	0.245	0.301	0.270	0.801	0.088

Figure 3. Word-level precision-recall curves on the source-disjoint test set.



4.3 Probability calibration

After Platt scaling, the MLP achieved the lowest ECE (0.0099) and NLL (0.1460), while the logistic model achieved the lowest Brier score (0.0366). The rule was worse on all three probability criteria. Its Platt slope of 0.145 indicates that the raw heuristic score needed strong compression; the MLP slope of 0.926 was much closer to a unit-scale mapping. Reliability curves confirm good average alignment at common low predicted probabilities but also reveal sparse high-confidence regions, which should not be interpreted from accuracy alone.

The calibration ranking does not exactly reproduce the thresholded F1 ranking, and that difference is substantively useful. The logistic model supplied the best selected operating point, whereas the MLP supplied the best AUROC, ECE, and NLL. A deployment concerned with catching as many hallucinated tokens as possible at one fixed review budget might prefer the logistic threshold. A system that changes its rejection level dynamically or constructs prediction sets may benefit more from the MLP's smoother ranking and probability alignment. Calibration therefore changes how model quality is used; it is not merely a cosmetic transformation of an already chosen binary output.

Table 6. Platt mappings and calibrated probability metrics on the test set.

Model	Platt slope	Intercept	Brier	ECE	NLL
Rule	0.145	-2.294	0.0402	0.0178	0.1738
Logistic	0.173	-1.444	0.0366	0.0117	0.1465
Compact MLP	0.926	-1.458	0.0369	0.0099	0.1460

Figure 4. Reliability diagrams for calibrated word-level hallucination probabilities on the test set.

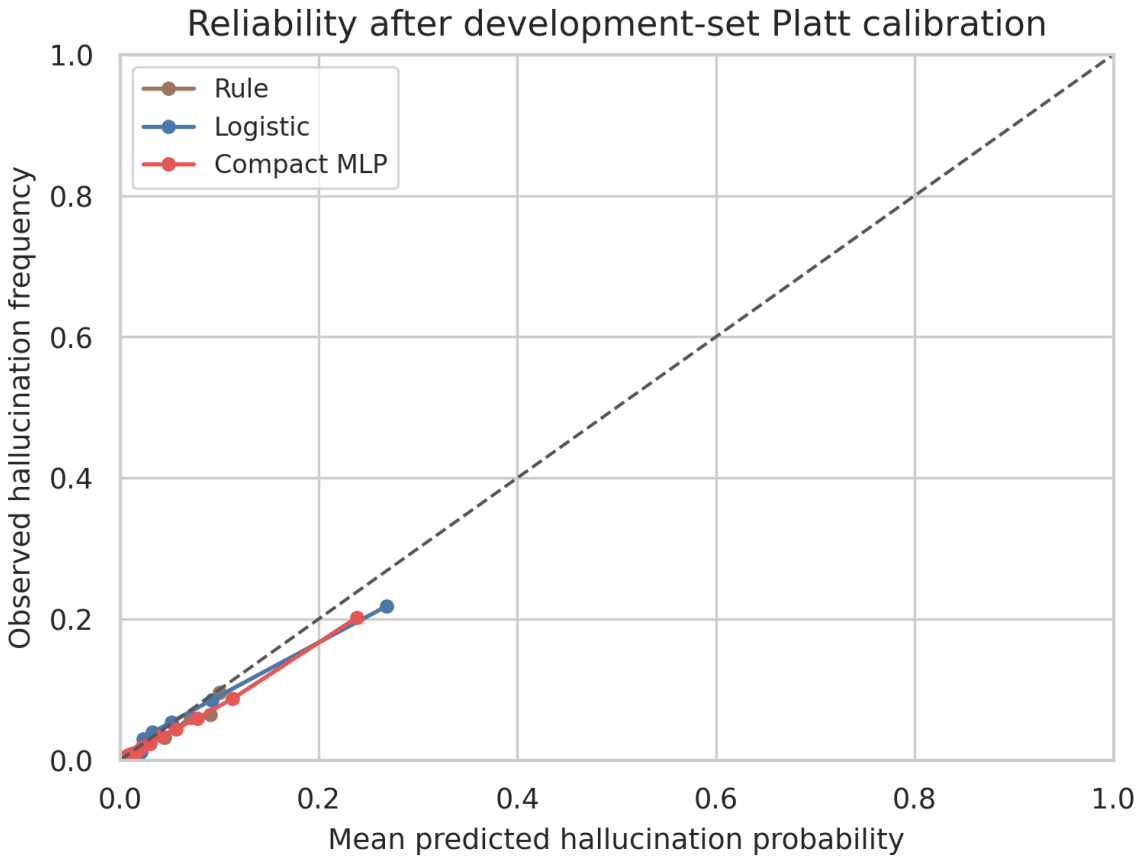
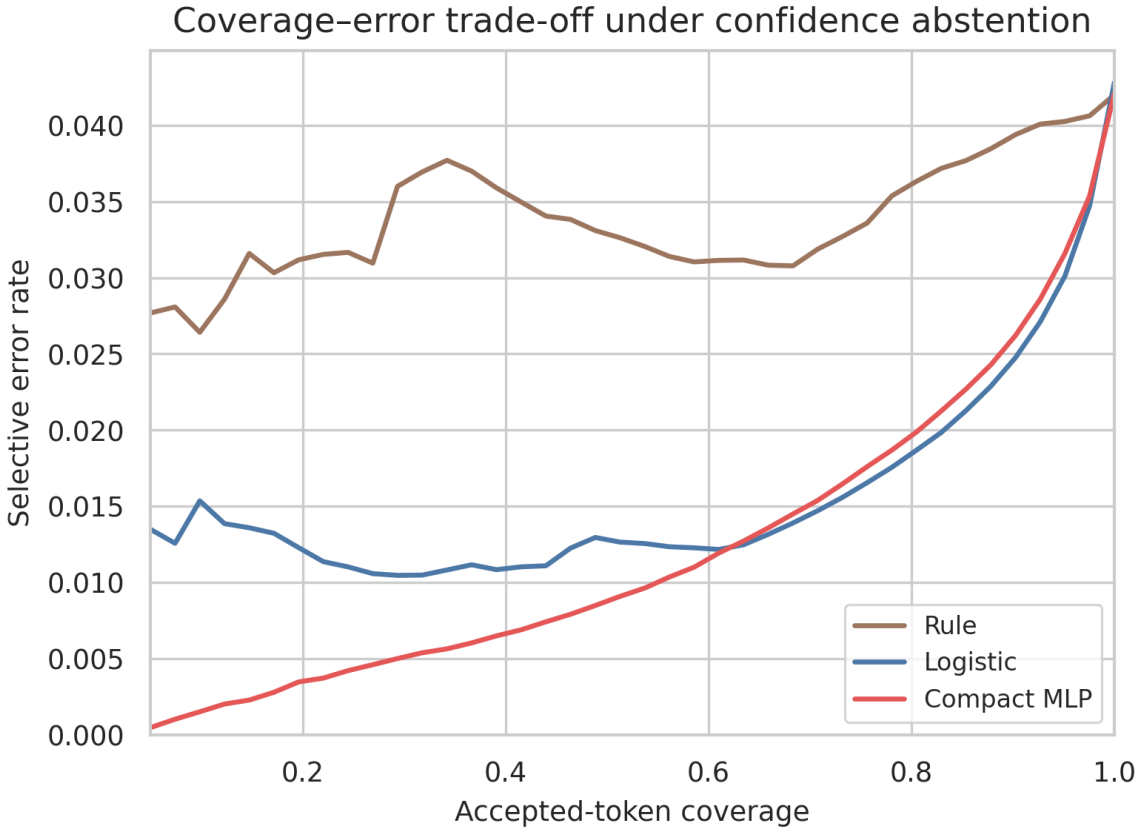


Figure 5. Selective error versus retained token coverage as confidence thresholds are swept.



4.4 Conformal coverage and efficiency

At 95% nominal coverage, pooled split conformal achieved 0.962 empirical coverage with mean set size 1.011. Task-Mondrian conformal achieved 0.960 with mean size 1.020. At 90% nominal coverage, pooled and Mondrian coverage were 0.917 and 0.918, with mean sizes 0.945 and 0.949; both methods occasionally returned an empty set, so mean size could fall below one. At 80% nominal coverage, the two methods reached 0.822 and 0.823. The observed values were above nominal in all nine pooled or aggregated task-Mondrian comparisons.

Coverage increased monotonically as the target became more conservative, but efficiency did not follow a single-label-only pattern. At the 95% target, both-label sets increased mean size above one; at lower nominal targets, empty sets became possible and pulled the mean below one. Empty and two-label sets have different operational meanings. An empty set signals that neither label satisfies the calibrated score threshold, whereas a two-label set records irresolvable uncertainty. A downstream workflow should route both cases for review, but it may prioritize them differently because one reflects score incompatibility and the other reflects ambiguity.

Task-level Mondrian results expose efficiency differences hidden by the aggregate. At 90% nominal coverage, QA obtained 0.934 coverage and a 0.982 singleton rate, while Summary and Data2txt obtained 0.912 coverage with singleton rates 0.933 and 0.937. At the 95% target, QA required a mean set size of 1.086 because 8.6% of sets contained both labels, whereas Summary had a mean size of 0.980 because 2.0% were empty. Group calibration therefore improved interpretability of task-specific behavior but did not force identical set-size profiles.

Source-block maximum calibration was fully covering but operationally uninformative. Its calibration quantile was 0.9789 at all three alpha values; every test set contained both labels, producing coverage 1.000, mean set size 2.000, singleton rate 0, and both-label rate 1.000. This is a genuine output of worst-token aggregation, not evidence that source-block calibration is uniformly superior. The construction protects against the most nonconforming token in a source and becomes extremely conservative when sources contain many heterogeneous tokens. Cluster-valid source-block sets should therefore be reported together with their efficiency cost.

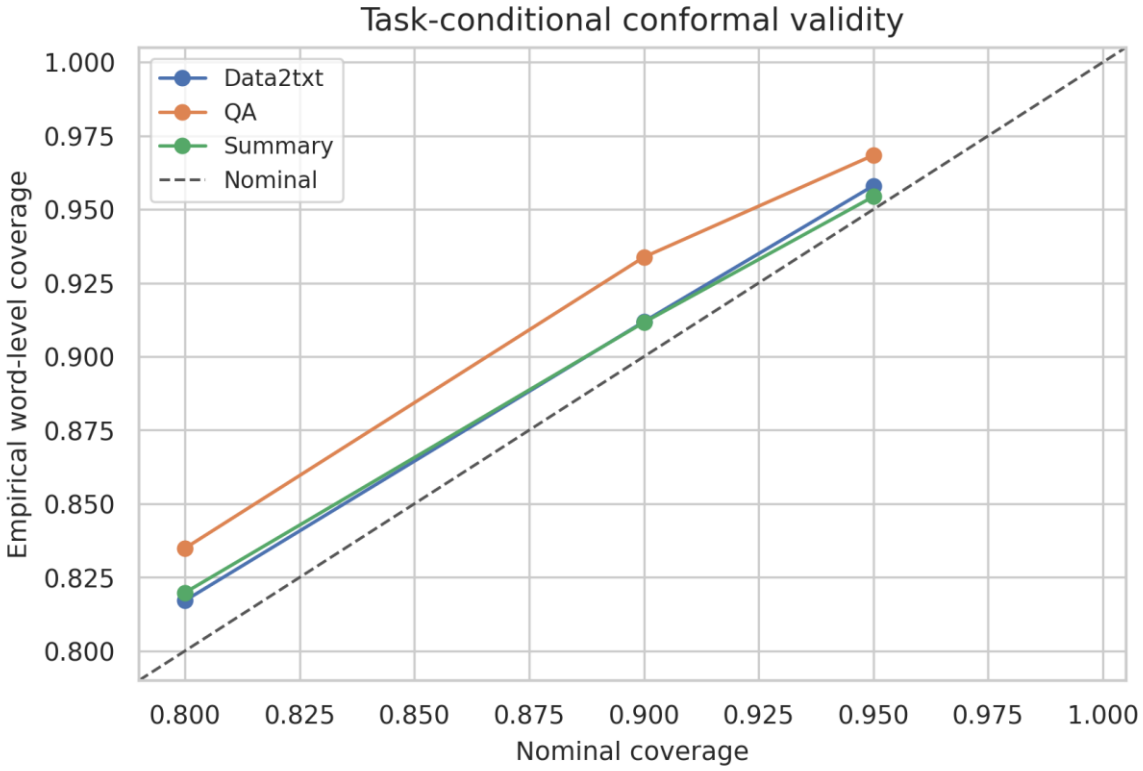
Table 7. Overall binary prediction-set performance at three nominal coverage levels.

Method	Nominal	Coverage	Mean size	Singleton	Both labels
Split-pooled	95.0%	0.962	1.011	98.9%	1.1%
Mondrian-task	95.0%	0.960	1.020	96.8%	2.6%
Source-block max	95.0%	1.000	2.000	0.0%	100.0%
Split-pooled	90.0%	0.917	0.945	94.5%	0.0%
Mondrian-task	90.0%	0.918	0.949	94.9%	0.0%
Source-block max	90.0%	1.000	2.000	0.0%	100.0%
Split-pooled	80.0%	0.822	0.840	84.0%	0.0%
Mondrian-task	80.0%	0.823	0.841	84.1%	0.0%
Source-block max	80.0%	1.000	2.000	0.0%	100.0%

Table 8. Task-Mondrian prediction-set performance by task and nominal level.

Group	Nominal	Coverage	Mean size	Singleton	Selective error
QA	95.0%	0.968	1.086	91.4%	0.035
Summary	95.0%	0.954	0.980	98.0%	0.027
Data2txt	95.0%	0.958	1.002	99.8%	0.042
QA	90.0%	0.934	0.982	98.2%	0.049
Summary	90.0%	0.912	0.933	93.3%	0.023
Data2txt	90.0%	0.912	0.937	93.7%	0.027
QA	80.0%	0.835	0.859	85.9%	0.028
Summary	80.0%	0.820	0.836	83.6%	0.019
Data2txt	80.0%	0.817	0.833	83.3%	0.019

Figure 6. Task-Mondrian empirical versus nominal word-level prediction-set coverage.



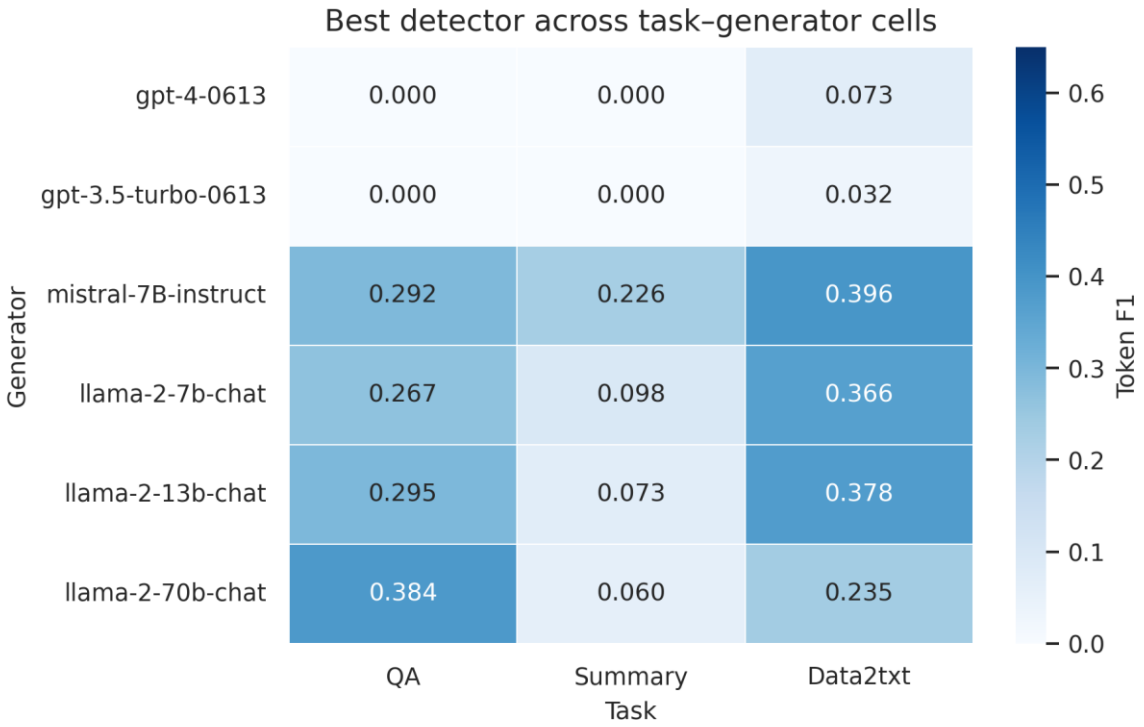
4.5 Task and generator shift

Performance also varied with the generating model. The logistic detector reached token F1 values from 0.021 on gpt-3.5-turbo-0613 and 0.049 on gpt-4-0613 to 0.319 on mistral-7B-instruct, 0.315 on llama-2-13b-chat, and 0.307 on llama-2-70b-chat. The two GPT subsets had positive-token rates of only 1.0% and 0.6%, so a fixed global threshold produced few true positives and unstable span estimates. The heatmap makes clear that task and generator composition jointly shape apparent detector quality; generator-level reporting is necessary when benchmark systems differ greatly in error prevalence.

Table 9. Generator-specific performance of the hashed logistic detector at the global threshold.

Generator	Positive rate	Precision	Recall	Token F1	AUPRC	AUROC
gpt-4-0613	0.6%	0.155	0.029	0.049	0.034	0.613
gpt-3.5-turbo-0613	1.0%	0.048	0.014	0.021	0.040	0.678
mistral-7B-instruct	7.1%	0.253	0.431	0.319	0.243	0.768
llama-2-7b-chat	6.2%	0.213	0.392	0.276	0.207	0.739
llama-2-13b-chat	5.4%	0.250	0.428	0.315	0.245	0.799
llama-2-70b-chat	5.0%	0.267	0.362	0.307	0.222	0.762

Figure 7. Token F1 across task-generator cells, showing the interaction of domain and generator.

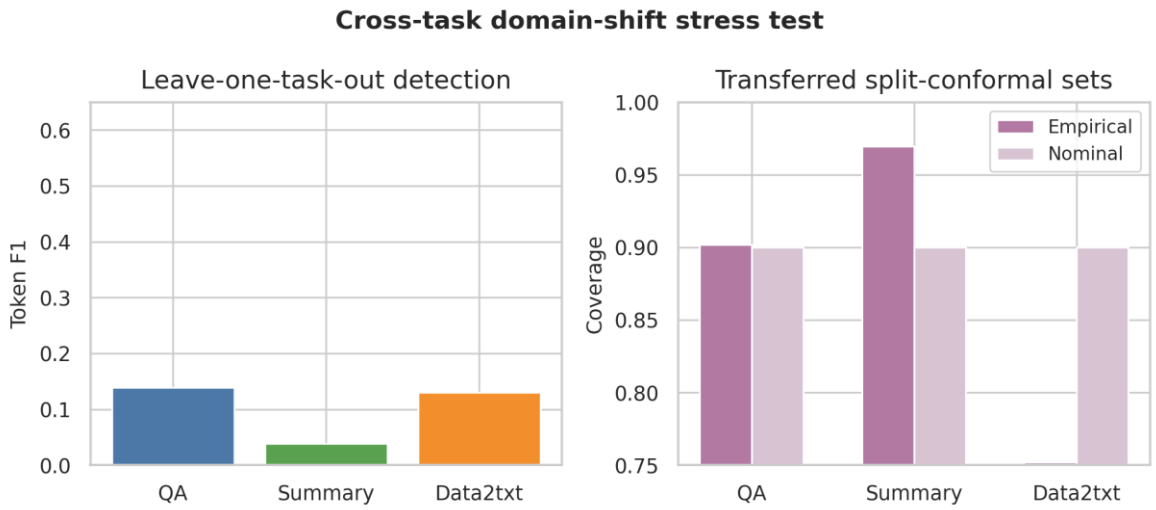


The LOTO test showed that ranking and coverage depended strongly on the held-out task. Holding out QA yielded token F1 0.139 and empirical 90% conformal coverage 0.902. Holding out Summary sharply reduced recall to 0.021 and F1 to 0.038, yet produced 0.970 coverage because the sets were almost always singleton and the positive class was sparse. Holding out Data2txt gave recall 0.510 but precision 0.074, F1 0.130, ECE 0.078, and only 0.752 conformal coverage. The last result is the clearest warning: standard calibration on QA plus Summary did not transfer its nominal coverage to structurally different Data2txt records.

Table 10. Leave-one-task-out discrimination, calibration, and empirical 90% prediction-set coverage.

Held-out task	Precision	Recall	Token F1	ECE	Coverage	Mean size
QA	0.196	0.108	0.139	0.021	0.902	0.945
Summary	0.272	0.021	0.038	0.017	0.970	0.999
Data2txt	0.074	0.510	0.130	0.078	0.752	0.778

Figure 8. Leave-one-task-out word-level F1 and empirical conformal coverage under task shift.



4.6 Ablation, computational cost, and synthesis

Removing evidence-overlap features reduced AUROC from 0.771 to 0.688 and token F1 from 0.210 to 0.125 in the controlled dense-feature ablation. Removing the local-context hash had little effect on F1 (0.211), whereas lexical-hash-only features increased recall to 0.645 but collapsed precision to 0.044. Evidence alignment was therefore the main source of discrimination, while contextual features controlled false positives.

Model fitting itself was inexpensive: hashed logistic training took 5.34 seconds, compact-MLP training took 4.49 seconds, and the shared test pass recorded 0.93 seconds. Feature construction dominated the run at 439.31 seconds, and total wall-clock runtime was 503.80 seconds. The distinction prevents model-only timings from understating the cost of tokenization, evidence alignment, dense and sparse feature extraction, calibration, subgroup evaluation, and figure generation. Even with that complete accounting, the end-to-end analysis finished in under nine minutes on the recorded environment, making repeated grouped calibration and shift checks feasible on a CPU-scale workflow.

Table 11. Controlled dense-feature ablation on the source-disjoint test set.

Configuration	Features	Precision	Recall	Token F1	AUPRC	AUROC
Full dense feature set	55	0.176	0.261	0.210	0.140	0.771
Without evidence-overlap features	41	0.076	0.364	0.125	0.070	0.688
Without local-context hash	39	0.173	0.270	0.211	0.140	0.769
Lexical hash only	16	0.044	0.645	0.083	0.049	0.530

Table 12. Measured model size and runtime from the completed experiment.

Model	Parameters / coefficients	Training (s)	Test batch (s)	Feature dimension
Rule	0	0.00	0.93	8
Hashed logistic	131,073	5.34	0.93	131,072
Compact MLP	2,337	4.49	0.93	55

Taken together, the results support three conclusions. First, a wide sparse linear representation was more effective at the selected threshold than the small nonlinear dense model, even though the MLP ranked and calibrated probabilities competitively. Second, calibrated prediction sets supplied a clearer uncertainty interface than a single threshold, but group choice determined efficiency. Third, exchangeable split results and LOTO shift results answer different questions. The former quantify finite-sample marginal behavior under the conformal assumptions; the latter measure observed transfer to a task

excluded from calibration. Weighted or adaptive shift methods could change that boundary, but the reported numbers belong to the unweighted executed protocol.

Several boundaries frame these conclusions. The labels assess support against the provided evidence, not the real-world truth of the evidence. Token dependence means that nominal token-level coverage should not be read as simultaneous coverage of an entire response. Generator comparisons are observational within the benchmark because generator identity is associated with task composition and error prevalence. Finally, LOTO provides only three coarse shifts. These limitations do not invalidate the measured comparisons; they define the population and prediction object to which the results apply.

5. Conclusions

This study evaluated word-level hallucination detection and selective prediction on all 17,790 RAGTruth responses with a source-disjoint fit, development, calibration, and test design. Among the three detectors, hashed logistic regression achieved the best thresholded token F1 (0.294), the compact 55-feature MLP achieved the best AUROC (0.813) and ECE (0.0099), and the deterministic overlap rule provided high recall at an unacceptable precision cost. Exact span agreement remained difficult, confirming that strong token ranking does not automatically yield clean span boundaries.

Pooled and task-Mondrian conformal sets met or exceeded their nominal aggregate targets at 95%, 90%, and 80% in the source-disjoint test. Source-block maximum calibration achieved complete coverage only by returning both labels for every token, making its conservatism explicit. The LOTO experiment then separated in-distribution calibration from cross-task behavior: coverage remained near nominal for held-out QA, became conservative for held-out Summary, and fell to 0.752 for held-out Data2txt. Accordingly, domain-shift coverage is reported as empirical, not as a distribution-free guarantee.

The practical lesson is that a RAG hallucination monitor should combine fine-grained localization, calibrated probabilities, and transparent selection diagnostics. Token F1, span overlap, calibration error, prediction-set coverage, set size, and risk-coverage curves each reveal a different failure mode. A source-disjoint protocol and explicit task-shift evaluation prevent optimistic leakage and overgeneralization, while the compact runtime keeps the complete analysis reproducible on modest hardware.

Funding: This research received no external funding.
Conflicts of Interest: The authors declare no conflict of interest.
ORCID iD (if any) Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers
Artificial Intelligence (AI) Use Disclosure: The authors declare that no artificial intelligence tools were used in the preparation of this manuscript.

References

[1] Angelopoulos, A. N., & Bates, S. (2023). Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning, 16*(4), 494–591. https://doi.org/10.1561/2200000101

[2] Angelopoulos, A. N., Bates, S., Fisch, A., Lei, L., & Schuster, T. (2024). Conformal risk control. In *The Twelfth International Conference on Learning Representations*. https://proceedings.iclr.cc/paper_files/paper/2024/hash/f3549ef9b5ff520a7e41ff3cc306ab2b-Abstract-Conference.html

[3] Bai, J., Chen, S., Zheng, D., & Kuo, M.-J. (2026). Interpretable attack-chain stage detection from AWS CloudTrail event sequences via linear models and HMM smoothing. *Information, Electrical and Electronics Engineering, 6*(1), 28–43. https://doi.org/10.33474/infotron.v6i1.24923

[4] Barber, R. F., Candès, E. J., Ramdas, A., & Tibshirani, R. J. (2023). Conformal prediction beyond exchangeability. *The Annals of Statistics, 51*(2), 816–845. https://doi.org/10.1214/23-AOS2276

[5] Bates, S., Angelopoulos, A. N., Lei, L., Malik, J., & Jordan, M. I. (2021). Distribution-free, risk-controlling prediction sets. *Journal of the ACM, 68*(6), Article 43, 1–34. https://doi.org/10.1145/3478535

[6] Campos, M., Farinhas, A., Zerva, C., Figueiredo, M. A. T., & Martins, A. F. T. (2024). Conformal prediction for natural language processing: A survey. *Transactions of the Association for Computational Linguistics, 12*, 1497–1516. https://doi.org/10.1162/tacl_a_00715

[7] Chang, X., Lu, Y., & Zhong, Z. S. (2026). Review-grounded explainable recommendation with faithfulness evaluation on Amazon reviews. *Journal of Electrical Engineering and Computer Sciences, 11*(1), 9–22. https://doi.org/10.54732/jeecs.v11i1.2

[8] Chen, S., He, S., & Sun, E. (2024). Risk-bounded GPU resource oversubscription via conformal demand envelopes in production AI clusters. *Journal of Advanced Computing Systems, 4*(5), 119–134. https://doi.org/10.69987/JACS.2024.40509

[9] Chen, Y., & Xu, H. (2026). Trust-calibrated multilingual RAG for humanitarian information platforms: Empirical evaluation on OMoS-QA for migration information access. *International Journal of Graphic Design, 4*(1), 141–164. https://doi.org/10.51903/ijgd.v4i1.3552

[10] Chen, Y., Zhang, Y., & Sherman, M. (2024). Going concern and bankruptcy prediction under extreme class imbalance: Cost-sensitive learning, resampling, and focal loss with explainable financial-ratio portraits. *Journal of Advanced Computing Systems, 4*(4), 80–96. https://doi.org/10.69987/JACS.2024.40407

[11] Chen, Y., Zhang, Y., Chau, D., & Sherman, M. (2023). Credit card default risk tiering with probability calibration and uncertainty-driven rejection: A reproducible study on the UCI credit card clients dataset. *Journal of Advanced Computing Systems, 3*(4), 31–47. https://doi.org/10.69987/JACS.2023.30403

[12] Chen, Y., Zhou, S., & Lin, E. (2025). Accounting-aware evidence retrieval for institutional due diligence of tokenized trade receivable RWA. *Journal of Technology Informatics and Engineering, 4*(3), 649–663. https://doi.org/10.51903/jtie.v4i3.542

[13] Desai, S., & Durrett, G. (2020). Calibration of pre-trained transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* (pp. 295–302). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.21

[14] Dziri, N., Kamalloo, E., Milton, S., Zaiane, O., Yu, M., Ponti, E. M., & Reddy, S. (2022). FaithDial: A faithful benchmark for information-seeking dialogue. *Transactions of the Association for Computational Linguistics, 10*, 1473–1490. https://doi.org/10.1162/tacl_a_00529

[15] El-Yaniv, R., & Wiener, Y. (2010). On the foundations of noise-free selective classification. *Journal of Machine Learning Research, 11*(53), 1605–1641. https://jmlr.org/papers/v11/el-yaniv10a.html

[16] Es, S., James, J., Espinosa Anke, L., & Schockaert, S. (2024). RAGAs: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations* (pp. 150–158). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.eacl-demo.16

[17] Farinhas, A., Zerva, C., Ulmer, D., & Martins, A. F. T. (2024). Non-exchangeable conformal risk control. In *The Twelfth International Conference on Learning Representations*. https://proceedings.iclr.cc/paper_files/paper/2024/hash/de04896f011beff76c91e094f72727f4-Abstract-Conference.html

[18] Gao, T., Yen, H., Yu, J., & Chen, D. (2023). Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 6465–6488). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.398

[19] Geifman, Y., & El-Yaniv, R. (2019). SelectiveNet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning* (pp. 2151–2159). PMLR. https://proceedings.mlr.press/v97/geifman19a.html

[20] Gibbs, I., & Candès, E. J. (2021). Adaptive conformal inference under distribution shift. *Advances in Neural Information Processing Systems, 34*, 1660–1672. https://proceedings.neurips.cc/paper/2021/hash/0d441de75945e5acbc865406fc9a2559-Abstract.html

[21] Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning* (pp. 1321–1330). PMLR. https://proceedings.mlr.press/v70/guo17a.html

[22] He, S., Chang, X., & Sun, E. (2024). Cross-cloud transfer learning for AI training capacity forecasting under workload and topology distribution shift. *Journal of Advanced Computing Systems, 4*(1), 100–120. https://doi.org/10.69987/JACS.2024.40108

[23] He, S., Li, C., & Rao, H. (2025). Few-shot cold-start workload forecasting for new AI inference tenants with time-series foundation models. *Journal of Technology Informatics and Engineering, 4*(1), 306–324. https://doi.org/10.51903/jtie.v4i1.546

[24] He, S., Tu, H., & Liu, I. (2023). Safe PD capacity forecasting with time-series foundation models and calibrated uncertainty for heterogeneous GPU clusters. *Journal of Advanced Computing Systems, 3*(4), 48–66. https://doi.org/10.69987/JACS.2023.30404

[25] Honovich, O., Aharoni, R., Herzig, J., Taitelbaum, H., Kukliansy, D., Cohen, V., Scialom, T., Szpektor, I., Hassidim, A., & Matias, Y. (2022). TRUE: Re-evaluating factual consistency evaluation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 3905–3920). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.naacl-main.287

[26] Jiang, Z., Araki, J., Ding, H., & Neubig, G. (2021). How can we know when language models know? On the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics, 9*, 962–977. https://doi.org/10.1162/tacl_a_00407

[27] Jin, J. (2025a). Evidence-chain reliable RAG: Hallucination detection, source attribution, and deterministic provenance explanations. *Journal of Technology Informatics and Engineering, 4*(2), 520–533. https://doi.org/10.51903/jtie.v4i2.535

[28] Jin, J. (2025b). LLM-style evidence cards for scientific search interfaces: A UI/UX design framework for retrieval transparency, ranking trust, and visual evidence hierarchy. *International Journal of Graphic Design, 3*(2), 397–414. https://doi.org/10.51903/ijgd.v3i2.3698

[29] Jin, J., Huang, T., & Lu, S. (2024a). Cost-sensitive learning, simulated PU learning, and one-class autoencoding for extreme-imbalance credit card fraud detection. *Journal of Advanced Computing Systems, 4*(6), 64–73. https://doi.org/10.69987/JACS.2024.40605

[30] Jin, J., Huang, T., & Lu, S. (2024b). A model-risk-friendly probability of default workflow: Calibration, distribution-free uncertainty quantification, and SHAP explanations on the UCI credit card default dataset. *Journal of Advanced Computing Systems, 4*(6), 74–85. https://doi.org/10.69987/JACS.2024.40606

[31] Kamath, A., Jia, R., & Liang, P. (2020). Selective question answering under domain shift. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5684–5696). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.503

[32] Kuo, M.-J., Zheng, D., & Hires, J. (2025). Federated topic-preference learning for knowledge-grounded chat with differential privacy. *Journal of Technology Informatics and Engineering, 4*(2), 385–401. https://doi.org/10.51903/jtie.v4i2.502

[33] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems, 33*, 9459–9474. https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html

[34] Li, C., Bai, J., & Wang, S. (2024). Evidence-chain reliable RAG: Word-level hallucination detection, source attribution, and provenance explanation for LLM applications. *Journal of Advanced Computing Systems, 4*(2), 76–92. https://doi.org/10.69987/JACS.2024.40207

[35] Li, C., Liu, G., & Zhao, Z. (2026). Cost-aware LLM-style routing for AIOps log analysis: Log parsing, anomaly detection, fault diagnosis, and incident summarization on LogEval task files. *Journal of Technology Informatics and Engineering, 5*(2), 91–103. https://doi.org/10.51903/jtie.v5i2.538

[36] Li, C., Zhou, B., & Gao, K. (2025). Risk-calibrated patient-facing AI safety cards: A UI/UX benchmark for explainable medical AI response interfaces. *International Journal of Graphic Design, 3*(2), 381–394. https://doi.org/10.51903/ijgd.v3i2.3709

[37] Li, J., & Zhou, A. (2026). Multi-regulation RAG for AI product counsel: A legal governance framework for cross-border digital commerces. *Rule of Law Studies Journal, 2*(2), 105–123. https://doi.org/10.64780/rolsj.v2i2.225

[38] Li, J., Cheng, X., Zhao, X., Nie, J.-Y., & Wen, J.-R. (2023). HaluEval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 6449–6464). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.397

[39] Li, Y. (2024). Findable then explainable: Retrieval–summary integration for code intelligence on a lightweight CodeSearchNet subset. *Journal of Advanced Computing Systems, 4*(7), 65–82. https://doi.org/10.69987/JACS.2024.40706

[40] Li, Y., & Lu, S. (2025). Language-guided feature selection for DDoS and intrusion detection on CICIDS2017. *Journal of Technology Informatics and Engineering, 4*(1), 284–305. https://doi.org/10.51903/jtie.v4i1.531

[41] Li, Y., Lu, S., & Zhao, L. (2025). LLM-as-design-critic: Aligning AI-generated UI feedback with human graphic design judgment. *International Journal of Graphic Design, 3*(1), 196–215. https://doi.org/10.51903/ijgd.v3i1.3661

[42] Li, Z., Zhang, K., & Wong, A. (2026). Numerical-reasoning guardrails for a quant research assistant: A compact reproducible benchmark using SEC and FRED data. *Journal of Technology Informatics and Engineering, 5*(2), 75–90. https://doi.org/10.51903/jtie.v5i2.541

[43] Li, Z., Zhou, S., & Zhou, Z. (2025). Financial risk dashboard design for institutional RWA investors: Visual hierarchy, chart comprehension, and explainability in FinChart-Bench. *International Journal of Graphic Design, 3*(1), 196–210. https://doi.org/10.51903/ijgd.v3i1.3715

[44] Liu, G., He, S., & Liu, I. (2023). LLM-augmented multi-source root cause attribution for CPU and network faults in microservices. *Journal of Advanced Computing Systems, 3*(6), 39–57. https://doi.org/10.69987/JACS.2023.30604

[45] Liu, G., He, S., & Wong, H. (2025). LLM-compatible visual brief cards for AI infrastructure capacity dashboards: A UI/UX framework for turning forecast risk into graphic design decisions. *International Journal of Graphic Design, 3*(1), 196–213. https://doi.org/10.51903/ijgd.v3i1.3723

[46] Liu, G., Li, C., & Zhang, E. (2024). OpsLLM for cloud incident triage: Bilingual RAG-based root cause analysis and alert summarization for AI infrastructure operations. *Journal of Advanced Computing Systems, 4*(4), 97–111. https://doi.org/10.69987/JACS.2024.40408

[47] Liu, T., Zhang, Y., Brockett, C., Mao, Y., Sui, Z., Chen, W., & Dolan, B. (2022). A token-level reference-free hallucination detection benchmark for free-form text generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 6723–6737). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.464

[48] Lu, S., & Zhou, D. (2024). TinyLLM-assisted intrusion detection for real-time IoT networks. *Journal of Advanced Computing Systems, 4*(8), 72–87. https://doi.org/10.69987/JACS.2024.40809

[49] Lu, S., & Zou, T. (2026). Uncertainty-aware medical vision–language classification on a lightweight MedMNIST-compatible biomedical patch benchmark. *Journal of Technology Informatics and Engineering, 5*(2), 1–19. https://doi.org/10.51903/jtie.v5i2.530

[50] Manakul, P., Liusie, A., & Gales, M. (2023). SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 9004–9017). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.557

[51] Meng, S., Chen, J., & Zheng, I. (2026). LLM-inspired offline reranking for financial search: Query rewriting, hybrid retrieval, and listwise relevance ranking on FiQA. *Journal of Technology Informatics and Engineering, 5*(1), 361–378. https://doi.org/10.51903/jtie.v5i1.537

[52] Mi, G., Ye, T., & Wood, D. (2025). A lightweight medical foundation model for cross-modal multi-task pretraining and parameter-efficient few-shot transfer on MedMNIST. *Journal of Technology Informatics and Engineering, 4*(3), 572–589. https://doi.org/10.51903/jtie.v4i3.492

[53] Min, S., Krishna, K., Lyu, X., Lewis, M., Yih, W.-t., Koh, P. W., Iyyer, M., Zettlemoyer, L., & Hajishirzi, H. (2023). FActScore: Fine-grained atomic evaluation of factual precision in long-form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 12076–12100). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.741

[54] Nie, J., & Zheng, D. (2024). Noisy-neighbor-aware VM degradation risk modeling with unsupervised residual fusion. *Journal of Advanced Computing Systems, 4*(4), 112–123. https://doi.org/10.69987/JACS.2024.40409

[55] Nie, J., Liu, G., Li, C., & Zou, T. (2026). Evidence-constrained incident visualization cards for distributed cloud logs: A UI/UX framework for turning Hadoop, OpenStack, and ZooKeeper logs into actionable SRE design interfaces. *International Journal of Graphic Design, 4*(1), 179–185. https://doi.org/10.51903/ijgd.v4i1.3703

[56] Niu, C., Wu, Y., Zhu, J., Xu, S., Shum, K., Zhong, R., Song, J., & Zhang, T. (2024). RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 10862–10878). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.585

[57] Particle Media. (n.d.). *RAGTruth* [Data set and source code]. GitHub. Retrieved July 29, 2026, from https://github.com/ParticleMedia/RAGTruth

[58] Quach, V., Fisch, A., Schuster, T., Yala, A., Sohn, J. H., Jaakkola, T. S., & Barzilay, R. (2024). Conformal language modeling. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=pzUhfQ74c5

[59] Rajpurkar, P., Jia, R., & Liang, P. (2018). Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 784–789). Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-2124

[60] Romano, Y., Sesia, M., & Candès, E. J. (2020). Classification with valid and adaptive coverage. *Advances in Neural Information Processing Systems, 33*, 3581–3591. https://proceedings.neurips.cc/paper/2020/hash/244edd7e85dc81602b7615cd705545f5-Abstract.html

[61] Ru, D., Qiu, L., Hu, X., Zhang, T., Shi, P., Chang, S., Cheng, J., Wang, C., Sun, S., Li, H., Zhang, Z., Wang, B., Jiang, J., He, T., Wang, Z., Liu, P., Zhang, Y., & Zhang, Z. (2024). RAGChecker: A fine-grained framework for diagnosing retrieval-augmented generation. *Advances in Neural Information Processing Systems, 37*. https://doi.org/10.52202/079017-0692

[62] Saad-Falcon, J., Khattab, O., Potts, C., & Zaharia, M. (2024). ARES: An automated evaluation framework for retrieval-augmented generation systems. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (pp. 338–354). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.naacl-long.20

[63] Shuster, K., Poff, S., Chen, M., Kiela, D., & Weston, J. (2021). Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021* (pp. 3784–3803). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-emnlp.320

[64] Su, W., Chen, S., & Qian, E. (2026). Narrative-aware scientific claim verification agent with evidence ranking for ClimateCheck. *Journal of Technology Informatics and Engineering, 5*(1), 327–340. https://doi.org/10.51903/jtie.v5i1.549

[65] Su, W., Chen, S., & Zhao, C. (2025). Budgeted multi-hop retrieval agent for compositional question answering: A retrieval-policy evaluation on the official MultiHop-RAG benchmark. *Journal of Technology Informatics and Engineering, 4*(3), 649–662. https://doi.org/10.51903/jtie.v4i3.543

[66] Su, W., Rao, H., & Ma, E. (2026). Privacy and data-integrity risk cards for LLM agents: A UI/UX design framework for secure human oversight under prompt-injection attacks. *International Journal of Graphic Design, 4*(1), 186–191. https://doi.org/10.51903/ijgd.v4i1.3699

[67] Sun, X., Lu, Y., & Chen, J. (2023). Controllable long-term user memory for multi-session dialogue: Confidence-gated writing, time-aware retrieval-augmented generation, and update/forgetting. *Journal of Advanced Computing Systems, 3*(8), 9–24. https://doi.org/10.69987/JACS.2023.30802

[68] Sun, X., Zhong, Z. S., & Wu, Q. (2026). Retrieval-grounded HDFS log anomaly detection and deterministic failure narrative generation. *Journal of Computational Systems and Applications, 3*(1), 15–30. https://doi.org/10.64229/j6d7fr94

[69] Tibshirani, R. J., Barber, R. F., Candès, E. J., & Ramdas, A. (2019). Conformal prediction under covariate shift. *Advances in Neural Information Processing Systems, 32*, 2530–2540. https://proceedings.neurips.cc/paper_files/paper/2019/hash/8fb21ee7a2207526da55a679f0332de2-Abstract.html

[70] Vovk, V. (2012). Conditional validity of inductive conformal predictors. In *Proceedings of the Asian Conference on Machine Learning* (pp. 475–490). PMLR. https://proceedings.mlr.press/v25/vovk12.html

[71] Wang, B., He, Y., Shui, Z., Xin, Q., & Lei, H. (2024). Predictive optimization of DDoS attack mitigation in distributed systems using machine learning. *Applied and Computational Engineering, 64*, 89–94. https://doi.org/10.54254/2755-2721/64/20241350

[72] Xin, Q. (2025a). Explaining OpenStack failure-injection log anomalies with retrieved normal prototypes. *Emerging Information Science and Technology, 6*(2), 125–146. https://doi.org/10.18196/eist.v6i2.31232

[73] Xin, Q. (2025b). Hybrid cloud architecture for efficient and cost-effective large language model deployment. *Journal of Information Systems and Informatics, 7*(3), 2182–2195. https://doi.org/10.51519/journalisi.v7i3.1170

[74] Xin, Q. (2025c). Uncertainty-aware late fusion for 3D perception (confidence calibration + fusion rule learning). *Journal of Technology Informatics and Engineering, 4*(1), 215–238. https://doi.org/10.51903/jtie.v4i1.485

[75] Xin, Q. (2026a). Auditable automated essay scoring and formative feedback: A rubric-grounded pipeline for secondary and higher education. *Journal of Applied Artificial Intelligence in Education, 2*(1), 1–19. https://doi.org/10.66053/jaaie.v2i1.348

[76] Xin, Q. (2026b). Early-warning analytics with LLM intervention rationales for student retention decisions: Classroom interaction modeling with xAPI-edu-data and dropout/success prediction. *Interdisciplinary Journal of Pedagogy and Research in Media Technology, 2*(1), 9–27. https://doi.org/10.64268/inspire.v2i1.117

[77] Xin, Q. (2026c). Explainable and fair credit risk scoring with counterfactual explanations: A reproducible evaluation on the German credit dataset (HELOC-motivated). *Journal of Information and Technology, 14*(2), 215–231. https://doi.org/10.32664/j-intech.v14i02.2228

[78] Xin, Q. (2026d). Host-based intrusion detection with system call sequences: Window localization and forensic narratives. *Aviation Electronics, Information Technology, Telecommunications, Electricals, and Controls, 8*(2), 325–334. https://doi.org/10.28989/avitec.v8i2.3973

[79] Xin, Q. (2026e). Log anomaly detection with conformal alert control and evidence-grounded incident ticket generation. *Aviation Electronics, Information Technology, Telecommunications, Electricals, and Controls, 8*(2), 247–264. https://doi.org/10.28989/avitec.v8i2.3974

[80] Xin, Q. (2026f). Probabilistic bike-sharing demand forecasting under changing weather and seasonal regimes with transformer-based models. *Findings*. https://doi.org/10.32866/001c.157499

[81] Xin, Q. (2026g). Self-supervised log anomaly detection with LogBERT-style transformers: Full empirical evaluation on a reproducible SynHDFS benchmark. *Journal of Electrical Engineering and Computer Sciences, 11*(1), 23–35. https://doi.org/10.54732/jeecs.v11i1.3

[82] Xin, Q., Xu, Z., Guo, L., Zhao, F., & Wu, B. (2024). IoT traffic classification and anomaly detection method based on deep autoencoders. *Applied and Computational Engineering, 69*, 64–70. https://doi.org/10.54254/2755-2721/69/20241511

[83] Xu, H., Chen, Y., & Med, A. (2025). Automatic detection and explanation of dark patterns from interface microcopy: Empirical comparison of BERT-style encoders, RoBERTa-style encoders, and LLM-style decoders on the ec-darkpattern dataset. *Journal of Technology Informatics and Engineering, 4*(3), 590–612. https://doi.org/10.51903/jtie.v4i3.491

[84] Zhang, B., Rao, H., & Zhao, D. (2024). Evidence-grounded RAG for cloud-native DevOps: Hallucination-resistant AIOps question answering over private operations documents. *Journal of Advanced Computing Systems, 4*(3), 109–125. https://doi.org/10.69987/JACS.2024.40308

[85] Zhang, B., Ren, Y., & Zou, J. (2025). LLM-style explainable e-commerce recommendation cards: A UI/UX design framework for trust-calibrated product recommendation. *International Journal of Graphic Design, 3*(2), 381–396. https://doi.org/10.51903/ijgd.v3i2.3697

[86] Zhang, B., Sun, X., Liu, G., & Zhou, B. (2026). LLM-style DevOps copilot for cloud-native troubleshooting: Retrieval-augmented runbook generation and command-safety evaluation. *Journal of Technology Informatics and Engineering, 5*(2), 104–118. https://doi.org/10.51903/jtie.v5i2.534

[87] Zhang, J. (2026). Early warning, grade prediction, and teacher-facing LLM-ready explanations toward an open volleyball course: Reproducible evidence from four public education datasets. *Journal of Technology Informatics and Engineering, 5*(2), 20–44. https://doi.org/10.51903/jtie.v5i2.525

[88] Zhang, K., Chen, Y., & Qian, A. (2025). Evidence-grounded accounting disclosure review cards: A visual communication framework for LLM-style explanations over SEC financial statements and notes. *International Journal of Graphic Design, 3*(2), 395. https://doi.org/10.51903/ijgd.v3i2.3710

[89] Zhang, R., Wen, Z., Wang, C., Tang, C., Xu, P., & Jiang, Y. (2025). Quality analysis and evaluation prediction of RAG retrieval based on machine learning algorithms. In *2025 5th International Conference on Computer Science, Electronic Information Engineering and Intelligent Control Technology (CEI)* (pp. 1012–1018). IEEE. https://doi.org/10.1109/CEI66465.2025.11398480

[90] Zhang, Y., & Zhang, H. (2025a). A therapist-facing session copilot for live counseling support: Reasoning-guided retrieval and ranking from multi-turn counseling dialogues. *Journal of Technology Informatics and Engineering, 4*(2), 464–486. https://doi.org/10.51903/jtie.v4i2.547

[91] Zhang, Y., & Zhang, H. (2025b). Visualizing the right counseling support: Evidence-linked recommendation cards for explainable mental health intake interfaces. *International Journal of Graphic Design, 3*(1), 214–229. https://doi.org/10.51903/ijgd.v3i1.3722

[92] Zhang, Y., & Zhou, Z. (2026). Strategy-aware therapist imitation for emotional support dialogues: A reproducible ESConv study for LLM response control. *Advances in Educational Technology and Psychology, 10*(2), 92–97. https://doi.org/10.23977/aetp.2026.100213

[93] Zhao, S., Bai, J., & Roberson, D. (2025). Multi-horizon GPU demand forecasting with workload semantics and operational risk curves: An empirical study on Alibaba Clusterdata GPU trace. *Journal of Technology Informatics and Engineering, 4*(3), 544–571. https://doi.org/10.51903/jtie.v4i3.498

[94] Zheng, D., & Li, C. (2024). Behavior-level jailbreak resistance via multi-stage refusal + utility preservation. *Journal of Advanced Computing Systems, 4*(1), 83–99. https://doi.org/10.69987/JACS.2024.40107

[95] Zheng, D., Li, C., & Davidson, H. (2023). Continual red-teaming for in-the-wild jailbreaks via online guardrail updates and guardrail distillation. *Journal of Advanced Computing Systems, 3*(2), 35–49. https://doi.org/10.69987/JACS.2023.30203

[96] Zheng, D., Zhang, B., & Geibel, J. (2024). VerifySafe: Toxicity-safe agent responses under adversarial prompts with evidence-based self-verification. *Journal of Advanced Computing Systems*, *4*(1), 67–82. https://doi.org/10.69987/JACS.2024.40106

[97] Zhong, Z. S., & Ling, S. (2024a). Improved theoretical guarantee for rank aggregation via spectral method. *Information and Inference: A Journal of the IMA, 13*(3), Article iaae020. https://doi.org/10.1093/imaiai/iaae020

[98] Zhong, Z. S., & Ling, S. (2024b). *Uncertainty quantification of spectral estimator and MLE for orthogonal group synchronization* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2408.05944

[99] Zhong, Z. S., Chen, J., Zhong, E., & Sun, X. (2025). Evidence-calibrated RAG for unanswerable question answering: Retrieval coverage, abstention calibration, and hallucination-proxy analysis on SQuAD 2.0. *Journal of Technology Informatics and Engineering, 4*(2), 502–520. https://doi.org/10.51903/jtie.v4i2.536

[100] Zhong, Z. S., Li, C., & Rao, H. (2026). Trajectory reliability prediction for generalist AI agents: Tool-use failure analysis and success forecasting on ZClawBench. *Journal of Technology Informatics and Engineering, 5*(1), 341–360. https://doi.org/10.51903/jtie.v5i1.539

[101] Zhong, Z. S., Pan, X., & Lei, Q. (2025). Bridging domains with approximately shared features. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics* (Vol. 258, pp. 559–567). PMLR. https://proceedings.mlr.press/v258/zhong25a.html

[102] Zhong, Z. S., Wu, Q., & Mi, G. (2025). Uncertainty-aware medical image explanation cards: LLM-generated visual explanations for AI-assisted radiology interfaces. *International Journal of Graphic Design, 3*(2), 415–436. https://doi.org/10.51903/ijgd.v3i2.3616

[103] Zhou, B., Jin, J., & Zhao, D. (2025). Calibrated resume-job matching for trustworthy LLM-assisted recruiter screening: Pairwise matching, probability calibration, and selective refusal on two public recruitment datasets. *Journal of Technology Informatics and Engineering, 4*(3), 625–648. https://doi.org/10.51903/jtie.v4i3.529

[104] Zhou, B., Li, C., & Liu, L. (2025). Risk-calibrated patient-facing AI safety cards: A UI/UX design framework for rubric-based medical risk communication. *International Journal of Graphic Design, 3*(2), 365–380. https://doi.org/10.51903/ijgd.v3i2.3696

[105] Zhou, S., Chen, Y., & Lee, K. (2026). Accounting-aware evidence-constrained agents for disclosure, settlement, and secondary-market risk monitoring in tokenized assets. *Journal of Technology Informatics and Engineering, 5*(2), 60–74. https://doi.org/10.51903/jtie.v5i2.544