

4th International Conference on Electronics, Engineering, Computer Science and Applied Development (EESD 2026)

Article

A Framework for Quantitative Evaluation and Causal Provenance Analysis of Hallucinations in Large Language Models

Hao-En Lee^{1,*}

¹ University of California, Irvine, Irvine, USA

* Correspondence: Hao-En Lee, University of California, Irvine, Irvine, USA

Abstract: Hallucination evaluation requires a clear distinction between unsupported content, detection errors, and the underlying mechanisms that generate an answer. This paper presents a comprehensive framework linking response-level measurement, annotation provenance, and controlled evidence perturbations. An executable instantiation analyzes the public RAGTruth corpus, comprising 17,790 archived responses associated with 2,965 distinct sources. Four lightweight detectors are rigorously evaluated using the official source-disjoint split, with uncertainty estimated by source-cluster bootstrap resampling. The corpus contains 7,664 responses with annotated hallucinations and 14,289 labeled spans. Excluding spans marked as implicitly true reduces the positive response count to 7,010, demonstrating that the annotation policy fundamentally changes the evaluation target. On 2,700 test responses, combined lexical and metadata logistic regression achieves an area under the receiver operating characteristic curve of 0.873, compared with 0.842 for metadata alone. Task-matched evidence replacement increases the lexical random forest score by 0.294 on average. Empty evidence instead decreases the lexical logistic regression score by 0.416, exposing a critical extrapolation failure. These perturbations identify the sensitivity of the detector to evidence while keeping responses fixed. They leave generation-level causal effects unestimated. Ultimately, the proposed framework provides an auditable evaluation procedure and establishes a practical boundary between measured detector behavior and hypotheses about hallucination origins.

Keywords: large language models; hallucination evaluation; causal provenance; evidence perturbation; retrieval augmentation

Received: 29 July 2026

Revised: 15 September 2026

Accepted: 28 September 2026

Published: 03 October 2026



Copyright: © 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Large language models produce fluent explanations, summaries, and answers from a mixture of supplied context and information learned during training. Their output can contain unsupported details even when the surrounding text is coherent. A reader may have difficulty locating these details because grammatical quality and evidential support are distinct properties [1]. Surveys of natural language generation describe hallucination across multiple tasks and distinguish measurement from mitigation. A recent survey of hallucination further organizes the problem around its taxonomy, contributing factors, detection methods, and limitations of retrieval-based systems. These distinctions become important when an evaluation is expected to explain why an error occurred.

An evaluation target must specify the reference against which an answer is judged. Faithfulness concerns agreement with an input source, whereas factuality concerns agreement with the world. Their separation is established in research on abstractive summarization [2, 3]. An answer can repeat an incorrect source faithfully, or introduce a

correct statement absent from that source. Treating both situations as the same binary outcome obscures the intended task. The problem becomes more difficult when a document is incomplete, a question requires external knowledge, or an annotation policy permits some reasonable inferences while rejecting others.

Public benchmarks expose different aspects of this problem. One benchmark evaluates whether models reproduce common falsehoods [4]. Another provides a benchmark for identifying hallucinated content in several generation settings. A third supplies archived responses and human-labeled hallucination spans for retrieval-augmented generation, supporting an analysis tied to explicit sources. A fourth evaluates long-form generations through atomic claims. These resources provide useful evidence, but their labels and sampling procedures define different estimands. A score from one benchmark cannot automatically be interpreted as a universal probability that a model is truthful.

Detection introduces a further layer. One method uses consistency across sampled generations for black-box hallucination detection. Another measures uncertainty at the level of meaning and targets confabulations, a subset of hallucinations. Such approaches address useful signals, yet disagreement among answers and absence of source support remain distinct observations [5]. A detector may also rely on superficial regularities, including answer length, task identity, or the model that produced the response. Consequently, an apparently strong aggregate score can coexist with weak source-sensitive reasoning within a particular task.

The explanatory question requires additional care. A feature associated with annotated hallucination is a candidate diagnostic signal. Changing that feature in a fitted detector reveals something about the detector. Establishing why the language model generated an unsupported claim requires either modifying the generation process or relying on justified assumptions connecting observed variables to that process. Causal inference distinguishes observation from modification through explicit structural assumptions. Without this distinction, a post hoc explanation of a classifier can be misreported as the origin of the text it classifies [6].

This study develops an evaluation and provenance framework around three linked questions. How does the measured hallucination burden change with task and annotation policy? How much diagnostic performance can be obtained from metadata without reading the evidence? How do fixed detectors respond when their evidence input is altered while the archived answer is preserved? The framework connects these questions through immutable response identifiers, source references, span coordinates, evaluation partitions, and modification records. Its empirical component uses public data and lightweight models whose features and settings can be reproduced without access to proprietary generation services.

The resulting contribution is a bounded audit of measurement and attribution [7]. It combines response prevalence with character-level annotation burden, contrasts lexical and metadata predictors, and estimates paired evidence-modification effects with source-level uncertainty. A structural representation separately identifies the additional observations required for generation-level claims. The experiments cover existing responses from six model configurations across three tasks. Their scope is the released corpus and the fitted detectors. They provide evidence about evaluation behavior, including a failure under empty context, while leaving internal language-model mechanisms and deployment performance for further investigation.

2. Related Work

2.1. Grounding and Retrieval Behavior

Retrieval-augmented generation integrates parametric language generation with information drawn from external collections. A formulation has been developed in which retrieved passages condition the generation process for knowledge-intensive tasks. Retrieval establishes an accessible evidence channel, enabling source attribution [8].

However, its presence does not guarantee that every generated claim is substantiated. A retrieved document may be irrelevant, the relevant passage may be missed during processing, or the response may incorporate information beyond the supplied material. These scenarios necessitate independent tracking of retrieval availability, generated content, and evaluation judgments.

Context access also varies with positional factors and task structure. Research has shown that models utilize relevant information unevenly across long inputs, with performance sensitive to the location of supporting material. Studies comparing parametric and nonparametric memory mechanisms have identified challenges in recalling less frequently encountered factual knowledge [9]. Both lines of work support methodological adjustments that preserve the query while varying evidence availability. In an archived-response setting, however, modifying the evidence presented to a detector cannot replicate the original generation experiments; instead, it assesses a downstream function operating on a fixed output.

Citation generation offers an alternative pathway to traceability. Systems that attach citations to generated text have been evaluated, with attention given to the alignment between claims and cited passages. A citation serves as a pointer to potential supporting material; its presence alone does not confirm that the claim logically follows from the cited document. Provenance records therefore require both the citation linkage and an explicit assessment criterion. The present framework implements this distinction at the dataset level: source identifiers remain separate from span annotations and detector scores, permitting independent inspection of each relationship [10].

2.2. Factual Consistency and Hallucination Detection

Question-based evaluation assesses whether information extracted from a generated text can be recovered from the source. Answers to questions derived from summaries serve as a signal of factual consistency. Research has explored improvements to question-answering-based evaluation and the interactions among components of the evaluation pipeline. These methods can reveal unsupported relations that lexical overlap fails to detect, although their performance depends on the quality of generated questions and extracted answers [11]. A failure in any auxiliary component may influence the final consistency score.

Natural language inference offers a complementary approach by estimating whether one text entails or contradicts another. Summary Consistency, or SummaC, addresses the mismatch between sentence-level inference and document-level consistency assessment through segmentation and aggregation. TRUE standardizes example-level evaluation across multiple factual consistency datasets and enables comparison of evaluation methods under a common protocol. These studies support reporting individual response decisions rather than relying solely on correlations between model-level averages. They also highlight the importance of the aggregation unit when multiple claims appear within a single response [12].

AlignScore learns a unified alignment function from multiple forms of alignment data. MiniCheck enables efficient verification of claims against grounding documents. These approaches represent stronger semantic alternatives to lexical probes [10]. They are reviewed as relevant comparison targets; however, their published results are not incorporated into the present experimental table because the models, data splits, preprocessing steps, and evaluation units differ and have not been harmonized. A future replication using semantic detectors could retain the same log and source-cluster evaluation framework while modifying only the scoring component.

The benchmarks discussed in the introduction differ in their methodological requirements. Claim-level factual precision necessitates claim decomposition and access to a knowledge source. Sampling-consistency methods require additional generations. Source-grounded span labels depend on a defined annotation policy and a stable mapping back to the original response [13]. RAGTruth is suitable for the current audit because its publicly available artifacts directly satisfy the mapping requirement. The experiment does

not infer missing token probabilities or reconstruct alternative generations from a single archived answer. It evaluates only the information present in the released output and explicitly identifies unobserved quantities.

2.3. Causal Analysis and Explanatory Scope

Modifications applied directly to internal model states can support a higher level of mechanistic explanation. Activation-based analyses have been used to identify computations underlying factual associations, and methods such as Rank-One Model Editing have been evaluated in this context. Such analyses rely on direct access to model states and a well-defined modification protocol. In contrast, the mere presence of a model name in an annotated corpus provides neither state access nor procedural specification. A provenance framework should accommodate internal modification evidence when available, without assuming that surface-level textual features correspond to the same computational mechanisms.

Causal representation learning focuses on the interplay among causal variables, learned representations, and shifts in data distributions. For hallucination analysis, relevant variables may include source adequacy, decoding constraints, intermediate representations, and evaluation criteria [14]. Some of these variables are only partially observable; others are not present in static archived datasets. Constructing a directed graph helps make modeling assumptions explicit, but graph structure alone does not establish causal effects. Effect identification depends on the generative process of variables, the feasibility of controlled variation, and the set of alternative explanations consistent with observed data.

Local Interpretable Model-agnostic Explanations (LIME) approximate classifier behavior using local surrogate models, while Shapley Additive Explanations (SHAP) distribute prediction influence across input features. These techniques operate on the fitted prediction function and its associated explanatory configuration. Their outputs support diagnostic tasks, yet attributing influence to a feature such as answer length does not imply that length caused the original generation error. Table 1 situates the present study among these approaches. Its empirical analysis operates on evidence supplied to the detector; its provenance representation maintains distinct fields for generation-level modifications and internal activation evidence [15].

Table 1. Evaluation and Explanation Approaches Distinguished by Observed Object and Intervention Scope.

Approach	Evaluation object	Evidence required
Source / claim evaluation	Support of a response or claim	Source text and assessment rule
Sampling uncertainty	Consistency across generations	Multiple sampled answers
Internal causal tracing	Model computation and predictions	Controlled activation or weight changes
Present empirical audit	Labels, detector scores, paired changes	Archived responses and modified detector evidence

3. Methodology

3.1. Evaluation Objects and Annotation Provenance

The framework represents an evaluation record as a source, an archived response, model metadata, and a set of labeled character intervals. A source identifier links the six responses generated from the same underlying input. Response identifiers distinguish model outputs without requiring the text to be unique. The original text remains unchanged during annotation auditing because whitespace normalization can invalidate interval coordinates. A separate deterministic representation is used for lexical feature

extraction [16]. Figure 1 shows the connection among immutable artifacts, detection, processing steps, and the final audit record.

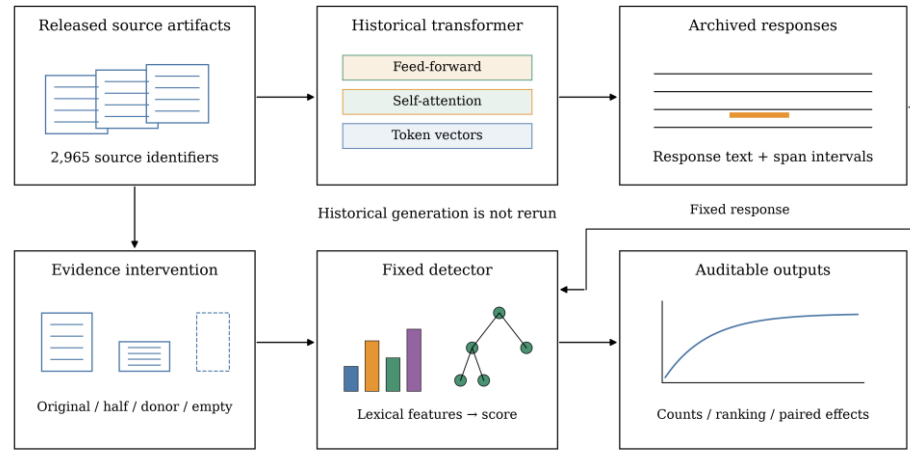


Figure 1. Implemented Audit Pipeline. The Transformer Sketch Denotes the Historical Generator; the Experiment Reads Its Archived Outputs and Does Not Execute This Block. Evidence Interventions Operate on the Detector Branch.

Let A_i denote the annotation set for response i , and let I_i contain its intervals marked implicitly true [17]. The released policy defines a positive response whenever A_i is nonempty. A sensitivity target removes I_i before assigning the response label. Both targets are given in Eq. (1), where the indicator is one when its condition holds. The sensitivity target is described as "excluding implicit truth." It remains a source-based annotation target and cannot be equated with a complete world-factuality assessment.

$$y_i = 1[A_i \neq \emptyset], \quad y_i^{\text{excl}} = 1[A_i \setminus I_i \neq \emptyset] \quad (1)$$

Response prevalence is the fraction of responses assigned a positive label. Character burden measures the total length of the union of labeled intervals divided by the total response length, as in Eq. (2). Here N is the number of responses in the evaluated collection, U_i is the union of half-open intervals belonging to A_i , and L_i is the number of Unicode characters in the original response. The union prevents double counting if intervals overlap. The denominator includes unmarked responses, allowing the metric to represent annotation burden across a collection [18]. It is a character-weighted quantity rather than an average of per-response percentages.

$$H = \frac{1}{N} \sum_{i=1}^N y_i, \quad B = \frac{\sum_{i=1}^N |U_i|}{\sum_{i=1}^N L_i} \quad (2)$$

An audit validates unique response identifiers, source references, source-level split consistency, interval bounds, and equality between each annotated substring and its recorded text. Missing annotation flags default to false. Evidence serialization retains the original summarization document, takes the passages field for question answering, and serializes structured data with sorted keys for data-to-text generation. Exact evidence hashes are checked across training and test sources [19]. This check addresses literal overlap under the chosen representation; paraphrases and related documents remain outside its coverage.

3.2. Diagnostic Representations and Fitted Models

Twelve lexical features characterize the response, its supporting evidence, and their surface-level relationship. Tokens are defined as alphabetic sequences with an optional internal apostrophe or digit sequences with an optional decimal part. Matching is performed on lowercase tokens. The features comprise logarithmic measures of response and evidence lengths, their length ratio, fractions of novel tokens and novel bigrams, fractions of novel numeric tokens and novel capitalized tokens, response lexical diversity, fractions of numeric and capitalized tokens, mean token length, and a logarithmic count

of punctuation-delimited sentences. Table 2 specifies how these features are grouped. Numerator counts are based on token occurrences, except for lexical diversity, which is computed over distinct tokens.

Table 2. Fixed Lexical Representation Used by Both Content-Based Detectors.

Feature group	Count	Definition
Lengths	3	$\log(1 + \text{response tokens}); \log(1 + \text{evidence tokens});$ response/evidence token ratio
Evidence novelty	4	Fractions of response tokens, bigrams, numbers, and capitalized tokens absent from evidence
Response composition	5	Unique-token fraction; number fraction; capitalized-token fraction; mean token length; $\log(1 + \text{punctuation groups})$

Every fractional feature uses a denominator of at least one; thus, a sequence containing no qualifying items yields a zero novelty fraction. The capitalization feature includes sentence-initial words and serves as a surface-level proxy rather than a named-entity recognizer. Bigrams consist exclusively of adjacent lowercase tokens. Numeric comparison is purely lexical: numerically equivalent values expressed via different strings are not considered matches. These definitions ensure computational executability while explicitly revealing representational constraints. For instance, a contradiction involving negation may retain high lexical overlap, whereas a semantically supported paraphrase may exhibit substantial lexical novelty [16].

Four detectors isolate content-based signals from corpus-level priors. Metadata logistic regression (LR) employs one-hot encoded task and generator identifiers. Lexical LR operates on the twelve lexical features. Combined LR concatenates both representations. All three models standardize input columns using training-set statistics and use an inverse regularization strength of one, a maximum of 2,000 optimization iterations, and random seed 42. Class distribution is preserved without reweighting. Equation (3) gives the logistic score, where z_i is the standardized input vector, w is the fitted coefficient vector, and b is the intercept [20].

Annotation spans, quality flags, source identifiers, and temperature values are excluded from all predictive inputs. Source identifiers serve only for record joining, partition validation, and clustered evaluation. Task and generator categories are encoded using training records, with previously unseen categories mapped to zero indicator values. These design choices prevent the detector from directly accessing target annotations or exploiting identifiers as memorization cues.

$$p_i = \frac{1}{1 + \exp(-(w^T z_i + b))} \quad (3)$$

The lexical random forest (RF) provides a nonlinear diagnostic probe based on the same twelve lexical features [11]. Following ensemble methodology, it averages class probabilities across 300 decision trees, each trained on bootstrap samples with a minimum leaf size of ten and square-root feature subsampling at each split. All trees use random seed 42. No hyperparameter tuning or test-set-driven model selection is performed. These models serve as diagnostic baselines, designed to expose differences between source-sensitive lexical features and metadata information under a consistent evaluation protocol.

Probability calibration is evaluated independently from ranking performance. Prior work indicates that learning algorithms can vary significantly in the reliability of their probability estimates, and calibration research emphasizes the importance of empirically assessing confidence rather than assuming it follows from classification accuracy. The Brier score in Eq. (4) computes the mean squared error between predicted probabilities and true binary labels across N held-out responses; lower values indicate better calibration. No post hoc calibration is applied, so probability outputs reflect each model's native behavior.

$$BS = \frac{1}{N} \sum_{i=1}^N (p_i - y_i)^2 \quad (4)$$

3.3. Controlled Evidence Interventions

The input to the detector is subject to controlled modification. Let f_θ denote a lexical detector with fixed parameters θ , R_i the archived response, and E_i the original evidence. The change in detector score under evidence transformation T_a is defined in Eq. (5). This transformation is applied prior to feature extraction [10]. The response text, fitted model parameters—including coefficients or tree structures—and the archived annotation record remain unchanged. Both lexical detectors are evaluated under identical transformations to isolate differences attributable solely to their scoring functions.

$$\Delta_i(a) = f_\theta(R_i, T_a(E_i)) - f_\theta(R_i, E_i) \quad (5)$$

Three evidence modifications are applied. First-half truncation retains the leading $\text{floor}(L/2)$ characters of the evidence string, where L denotes its total character length. Task-matched replacement substitutes the evidence with the complete evidence from another test source assigned to the same task; a seeded permutation followed by a cyclic shift ensures a derangement—no source receives its own evidence [21]. All responses associated with a given source receive the identical replacement evidence. Empty evidence substitution replaces the entire evidence string with an empty string. The donor mapping is determined by seed 2026 and remains fixed across all reported detector evaluations.

The mean paired effect and decision-flip fraction are defined in Eq. (6), where $p_i(a)$ denotes the transformed score and p_i the original score. A fixed decision threshold of 0.5 is used throughout the study. The mean absolute score change is also reported, as signed changes may offset one another in the arithmetic mean. These metrics characterize the behavior of the fitted detector on this test collection. Notably, the confidence interval for task-matched replacement is conditional on the specific donor mapping employed; it does not marginalize over all possible assignment configurations.

$$\begin{aligned} \delta(a) &= \frac{1}{N} \sum_{i=1}^N \Delta_i(a) \\ F(a) &= \frac{1}{N} \sum_{i=1}^N 1[1[p_i(a) \geq 0.5] \neq 1[p_i \geq 0.5]] \end{aligned} \quad (6)$$

Modifying the evidence may alter which statements are supported by the input. The archived gold label corresponds exclusively to the original response–evidence pair. Therefore, transformed scores are compared directly to their original counterparts, and classification accuracy under transformed evidence is not assessed against the original labels. Computing such accuracy would necessitate new human annotations for every modified response–evidence pair. Empty evidence further serves as an out-of-distribution probe: it evaluates detector behavior when evidence is absent, despite the training data always providing evidence.

3.4. Causal Provenance and Uncertainty

Figure 2 separates historical generation from subsequent evaluation. The source and task influence the prompt; the prompt, generator configuration, and decoding conditions influence the response. Annotation policy, response, and source jointly determine the released label. The detector reads response and evidence, with metadata added in the combined model [22]. This directed acyclic graph is a structural account of the workflow. Unobserved source difficulty, training exposure, and annotator judgment are acknowledged rather than estimated from proxy features.

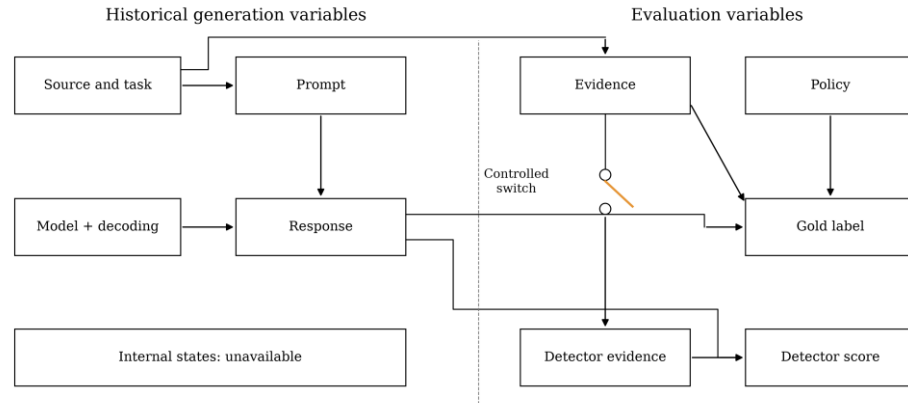


Figure 2. Causal Scope of the Analysis. Solid Arrows Describe the Recorded Workflow. The Controlled Switch Changes Detector Evidence While the Response Is Fixed. Generator-state Interventions Are Specified Conceptually and Are Not Executed.

For a generation-level study, Eq. (7) defines a separate potential-outcome contrast between evidence condition a and reference condition a_0 . $Y_i^{\text{gen}}(a)$ is the label assigned to a newly generated response under condition a , evaluated using a prespecified, common reference policy. The expectation would average over a defined prompt population and generation randomness. Identifying this contrast requires controlled generation, stable evaluation criteria, and recorded decoding conditions. The present archive has no paired regenerated outcomes under the alternative conditions, so this estimand is left unestimated. Observed differences between named generators remain descriptive comparisons.

$$\tau^{\text{gen}}(a, a_0) = E[Y_i^{\text{gen}}(a) - Y_i^{\text{gen}}(a_0)] \quad (7)$$

Held-out ranking uses area under the receiver operating characteristic curve (AUROC) and average precision (AP). Precision–recall analysis is useful when positive prevalence varies; AP is reported as the step-weighted summary of the precision–recall curve. Macro F1 averages the positive and negative class F1 scores. Balanced accuracy averages sensitivity and specificity. Confusion counts accompany threshold-based interpretation. All these quantities use the original evidence and the released response labels unless a sensitivity analysis explicitly changes the target.

Uncertainty is estimated through bootstrap resampling [21]. The resampling unit is the source, preserving its six linked responses. Each of 1,000 replicates samples 450 test sources with replacement, concatenates all their responses, and recomputes the statistic. Equation (8) gives the 95% confidence interval (CI), where q_α denotes the alpha quantile of the replicate statistic values. The same resamples are used for paired model comparisons and condition contrasts. Models are not refitted within bootstrap replicates, so intervals describe held-out sampling uncertainty conditional on fitted models.

$$CI_{95\%} = [q_{0.025}, q_{0.975}] \quad (8)$$

The official partition is retained after checking for exact cross-split evidence duplicates. Training transformations are fitted on training records only, and the test set is used for evaluation. These choices address common leakage concerns in machine-learning-based science. The reproducibility record retains input hashes, preprocessing rules, software versions, model settings, detector predictions, and condition specifications. Excluding implicitly true spans and restricting evaluation to good-quality responses are target and sample sensitivity analyses, respectively; neither analysis retrains the detectors.

4. Experiments and Results

4.1. Public Corpus and Audit Results

The experiment uses the public RAGTruth corpus released in association with a formal conference publication. The `source_info.jsonl` and `response.jsonl` files were retrieved from the official repository on September 14, 2026. They contain 2,965 sources

and 17,790 responses, with exactly six responses per source [23]. The release covers summarization, question answering, and data-to-text generation. The summarization sources comprise news articles from major international outlets and recent news reports; the remaining source categories are drawn from established machine reading comprehension benchmarks and publicly available review datasets. Table 3 presents counts computed directly from the released records.

Table 3. Corpus Statistics Computed from the Released Files. Character Burden Uses the Union of Annotated Intervals and All Response Characters within Each Task.

Task	Sources	Responses	Positive	Spans	Burden (%)
Data-to-text	1033	6,198	4,254	9,290	4.79
Question answering	989	5,934	1,724	2,927	8.23
Summarization	943	5,658	1,686	2,072	3.07
Total	2,965	17,790	7,664	14,289	5.30

The official partition contains 15,090 training responses from 2,515 sources and 2,700 test responses from 450 sources. Each test task contributes 900 responses from 150 sources. Every response identifier is unique, every source reference resolves correctly, and no source identifier crosses the partition boundary. The exact-evidence audit identifies one duplicate context group within the corpus and confirms no group spans both training and test sets. No training record is removed. This result supports the source-disjoint evaluation protocol used here, while semantic near-duplicate overlap remains unmeasured.

All 14,289 labeled spans have valid half-open boundaries, and every extracted substring matches its recorded annotation text exactly [24]. The release classifies 17,617 responses as good, 144 as incorrect refusals, and 29 as truncated. The primary analysis retains all records to preserve the official evaluation population. A separate high-quality test subset contains 2,675 responses. Refusal quality and hallucination are treated as distinct attributes: an unhelpful refusal may be unmarked for hallucination without constituting a satisfactory answer to the original request.

The retained input hashes begin e4c2e4ac24fff676 for response.jsonl and 0dffc26ea9f3c1c3 for source_info.jsonl; complete hashes are stored with the executable analysis. Computation uses Python 3.12.14, NumPy 2.5.2, and scikit-learn 1.9.0. No additional language-model responses, token probabilities, embeddings, or human labels are generated. The repository supplies a permissive software license; underlying source material should be reused in accordance with its original licensing terms. The manuscript reports aggregate statistics rather than redistributing source articles or personal review text.

4.2. Quantitative Hallucination Profiles

Across the complete corpus, 7,664 responses contain at least one labeled hallucination span, representing 43.08% of all responses. The character-weighted hallucination burden is 5.30%. These two metrics address distinct aspects: response prevalence reflects how frequently an answer includes any annotated hallucinatory content, whereas character burden quantifies the proportion of total generated text covered by such annotations. A long response with a single brief unsupported phrase counts as positive for prevalence but contributes minimally to the character-weighted total. Neither metric captures the functional significance or potential impact of the unsupported claim.

Data-to-text generation exhibits the highest response prevalence, with 4,254 positive cases among 6,198 responses (68.64%), compared to 29.05% for question answering and 29.80% for summarization. However, the character-weighted burden follows a different ranking: question answering reaches 8.23%, data-to-text generation 4.79%, and summarization 3.07%. Figure 3 illustrates this divergence. The combination of high prevalence and relatively low character burden in data-to-text generation demonstrates that the two aggregation methods yield different task rankings. This does not imply lesser

practical impact, as annotation does not encode the functional importance of individual claims.

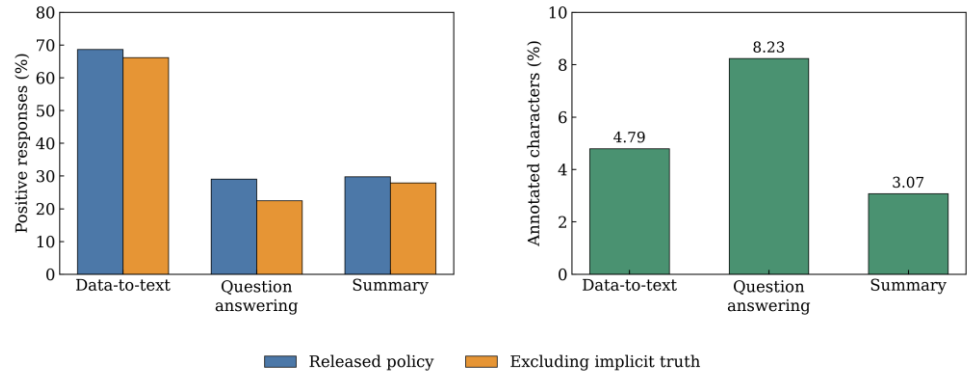


Figure 3. Task-level Prevalence under Two Annotation Policies and Character-Weighted Burden. The Corpus-Wide Analysis Uses All 17,790 Responses.

The span labels consist of 5,324 clearly unsupported factual assertions, 6,237 clearly unsubstantiated statements, 2,527 nuanced unsubstantiated statements, and 201 nuanced unsupported factual assertions. Additionally, 1,928 spans are marked as implicitly verifiable and 1,642 as associated with null source fields. These flags indicate annotation context and are not interpreted as experimentally confirmed causes. Excluding implicitly verifiable spans reduces the count of positive responses from 7,664 to 7,010—a decrease of 654 responses, or 3.68 percentage points—yielding a revised prevalence of 39.40%. Some responses retain other annotated spans and therefore remain classified as positive.

The effect of excluding implicitly verifiable spans is most pronounced in question answering, where the positive count declines from 1,724 to 1,333, shifting prevalence from 29.05% to 22.46%. For data-to-text generation, the count drops from 4,254 to 4,101; for summarization, from 1,686 to 1,576. This variation underscores the importance of explicitly stating the annotation policy used in evaluation reports [21]. The analysis adopts the labeling scheme provided by the dataset authors and does not independently verify the factual accuracy of flagged statements against external sources.

4.3. Held-Out Detection and Metadata Dependence

The test set contains 943 positive instances among 2,700 responses, yielding a prevalence of 34.93%. Table 4 compares the performance of four detection methods under the original annotation policy. The combined logistic regression model achieves an AUROC of 0.873 with a source-bootstrap confidence interval of 0.858–0.886 and an average precision (AP) of 0.801. The metadata-only logistic regression attains an AUROC of 0.842 and AP of 0.743 without access to the response or source text [25]. The lexical random forest model reaches an AUROC of 0.831 and AP of 0.741. The lexical logistic regression model yields an AUROC of 0.804 and AP of 0.674.

Table 4. Official Held-Out Response-Level Detection. Auroc Intervals Use 1,000 Source-Cluster Bootstrap Replicates; Other Entries Are Point Estimates. Lower Brier Score Is Preferable.

Detector	AUROC (95% CI)	AP	Macro F1	Balanced accuracy	Brier
Metadata LR	0.842 (0.825–0.858)	0.743	0.755	0.752	0.155
Lexical LR	0.804 (0.788–0.819)	0.674	0.723	0.728	0.173
Combined LR	0.873 (0.858–0.886)	0.801	0.782	0.783	0.141
Lexical RF	0.831 (0.816–0.846)	0.741	0.751	0.750	0.160

The combined model improves AP over the lexical logistic regression model by 0.127; the paired source-bootstrap confidence interval for this difference is 0.101–0.152. Its Brier score is 0.141, compared with 0.155 for metadata logistic regression, 0.160 for lexical random forest, and 0.173 for lexical logistic regression. Figure 4 displays the receiver

operating characteristic and precision–recall curves. The results indicate that metadata contain substantial predictive signal in this corpus [26]. However, a metadata-only score cannot supply claim-specific evidence, even when its overall ranking performance remains competitive.

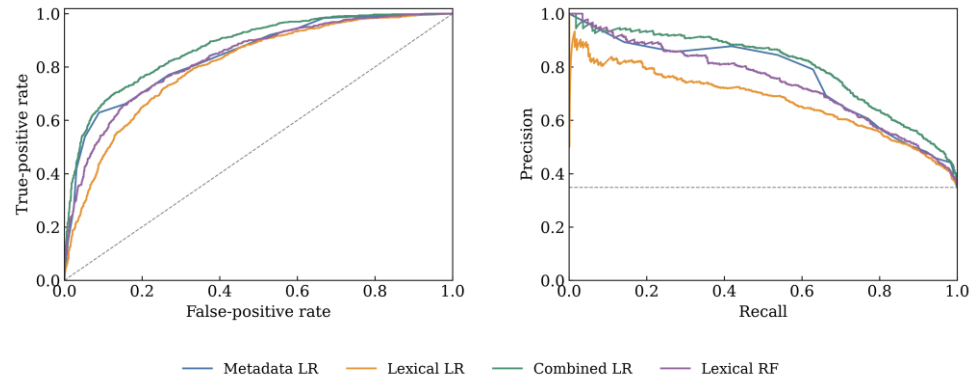


Figure 4. Held-out Receiver Operating Characteristic and Precision–recall Curves for Four Diagnostic Baselines. The Dashed Precision Reference Equals Test Prevalence.

At a fixed threshold of 0.5, the combined logistic regression model identifies 683 true positives, 260 false negatives, 277 false positives, and 1,480 true negatives. Its macro F1 score is 0.782 and balanced accuracy is 0.783. Thus, even the highest-performing method among these baselines fails to flag 260 of the 943 annotated positive responses [27]. This threshold represents a fixed diagnostic operating point; operational deployment would require independent evaluation of error costs, human review capacity, and potential consequences of misclassifying factually supported statements.

Task-specific analysis alters the interpretation of aggregate metrics. For data-to-text generation, lexical logistic regression achieves a balanced accuracy of 0.510 despite its pooled AUROC of 0.804, reflecting a strong bias toward predicting positive labels in that task. For summarization, its balanced accuracy is 0.528. Lexical random forest achieves the highest average precision among the four methods on question-answering, with an AP of 0.547 compared to 0.530 for the combined model [28]. Figure 5 presents task-specific average precision across the 900 test responses per task, using the released annotation policy and original evidence. These variations demonstrate that pooled metrics are influenced by task composition and cannot replace within-task evaluation.

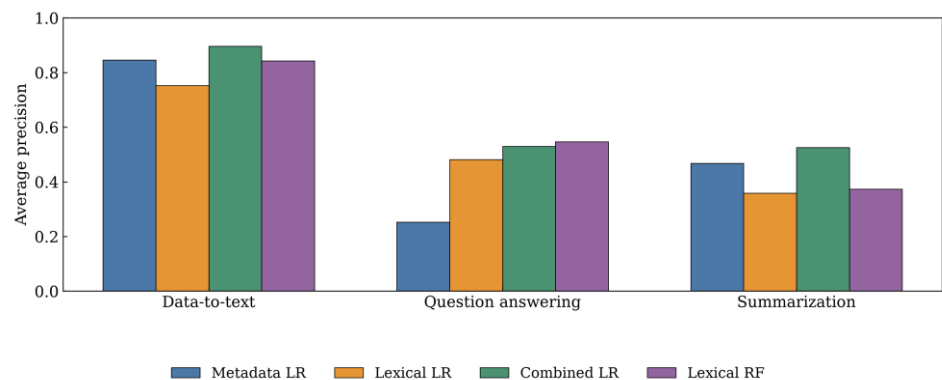


Figure 5. Task-specific Average Precision on the 900 Test Responses Per Task, Using the Released Annotation Policy and Original Evidence.

Prevalence of positive responses varies across generator configurations, ranging from 13.52% for gpt-3.5-turbo-0613 to 65.87% for mistral-7B-instruct across the full corpus. These labels reflect historical configuration identifiers in the dataset release. The

temperature parameter is set to 0.7 for both named Generative Pre-trained Transformer (GPT) configurations, while the other four configurations include multiple values between 0.7 and 1.0. Consequently, architecture, training methodology, decoding strategy, and output behavior co-vary. The observed differences do not isolate individual factors such as parameter count, specific training approaches, or temperature settings, and should not be interpreted as a comparative ranking of contemporary model services.

4.4. Evidence Interventions and Failure Analysis

Table 5 reports score changes for the two lexical detectors under the three evidence modifications. First-half truncation increases the logistic score by 0.174 on average, with a source-bootstrap interval of 0.167–0.182. The random forest increase is 0.165, with an interval of 0.156–0.175. The corresponding decision-flip fractions are 24.48% and 30.19%. These changes arise from altering evidence presented to a fixed scorer. They combine lost content with changes in evidence length and overlap statistics.

Table 5. Paired Detector-Input Interventions on All 2,700 Test Responses. Signed and Absolute Changes Are in Probability Units. Flip Rates Use the Fixed Threshold 0.5.

Evidence	Detector	Mean change (95% CI)	Mean absolute change	Flip (%)
First half	Lexical LR	+0.174 (+0.167, +0.182)	0.175	24.48
First half	Lexical RF	+0.165 (+0.156, +0.175)	0.178	30.19
Shuffled	Lexical LR	+0.433 (+0.413, +0.452)	0.434	60.30
Shuffled	Lexical RF	+0.294 (+0.277, +0.310)	0.301	51.89
Empty	Lexical LR	-0.416 (-0.436, -0.396)	0.416	38.26
Empty	Lexical RF	+0.216 (+0.199, +0.232)	0.252	47.93

Task-matched replacement increases the logistic score by 0.433 and the random forest score by 0.294. Their intervals are 0.413–0.452 and 0.277–0.310, respectively. Decision flips occur for 60.30% and 51.89% of responses. The mean absolute changes, 0.434 and 0.301, are close to the signed changes, indicating that increases dominate these aggregate contrasts [23]. The modification changes the entire evidence string, including topic and lexical composition. It does not identify the individual sentence responsible for a particular score increase (As shown in Figure 6).



Figure 6. Mean Detector-Score Changes under Fixed-Response Evidence Interventions. Error Bars Show Conditional 95% Source-Bootstrap Intervals; the Zero Line Indicates No Score Change.

Empty evidence produces divergent behavior. The random forest score increases by 0.216, whereas the logistic score decreases by 0.416, with an interval of −0.436 to −0.396. For the linear detector, removal of the source can therefore make an answer appear less likely to contain an annotated hallucination. The denominator floor prevents numerical division by zero but does not make the resulting feature vector representative of training data. The response-to-evidence length ratio ranges from 0.0057 to 2.7281 during training and from 7 to 361 under empty evidence. Its mean contribution to the linear predictor

changes by -50.686 logit units, outweighing the positive contribution of token novelty. This coefficient-level decomposition identifies the dominant pathway for the empty-context failure within the fitted detector.

The empty-context result motivates inclusion of an explicit evidence-availability state in a practical evaluator. An unavailable source should produce a missing-evidence outcome before a support score is interpreted. The study records this as an implementation requirement; the modification table remains the output of the unmodified probes. Neither the decrease for the linear model nor the increase for the forest establishes the correctness of transformed labels. A semantic reassessment of the modified response-evidence pairs would be necessary to evaluate that question.

4.5. Sensitivity and Reproducibility Boundaries

Restricting evaluation to high-quality test responses changes the combined-model AUROC from 0.873 to 0.872 and AP from 0.801 to 0.801 after rounding. Excluding implicitly true spans reduces the number of test positives from 943 to 894. With detector predictions held fixed, the combined-model AUROC remains 0.873 and AP becomes 0.790 under this alternative definition. The near-identical AUROC values do not alter the underlying definition of an error; rather, they indicate that the ranking of these fixed predictions is largely unaffected by the composition of the flagged subset.

The calculations rely on a single official data split, fixed random seeds for model fitting, and one task-aligned donor mapping. Bootstrap confidence intervals preserve source cluster structure but do not account for variability across training seeds, annotation consistency, or donor selection uncertainty. Exact hash verification ensures computational reproducibility but does not guarantee semantic equivalence; no external benchmark is used to assess generalization performance. All saved predictions are provided to support independent re-evaluation of the reported metrics. Enhancing semantic detection capability and applying repeated generation adjustments represent logical next steps, though their effectiveness and underlying mechanisms are not established by the present lexical-level analysis.

5. Discussion

The results suggest that hallucination auditing benefits from maintaining three distinct components: the annotation target, the detector output, and the record of model response modification. Response prevalence and character burden exhibit different task sequencing patterns, while the handling of implicitly true material alters the composition of the positive set. A single aggregate metric obscures these distinctions. Documenting the reference policy enhances interpretability for subsequent comparisons and enables clear separation between changes in evaluation criteria and changes in model behavior.

Metadata dependence introduces a practical tradeoff. Generator and task identifiers improve pooled prediction performance on corpora with known configuration distributions, potentially aiding prioritization of manual review within such controlled settings. However, a metadata-only model provides no direct evidence regarding individual claims, and its informative priors may shift when new generators or tasks are introduced. A provenance report should explicitly indicate which inputs contribute to its score and preserve a content-based comparison under identical data splits [29, 30].

The empty-evidence experiment demonstrates that intervention magnitude does not inherently reflect quality. A detector may respond strongly to an input transformation due to feature-space artifacts; another may respond in a plausible direction without verifying semantic grounding. Handling of missing evidence, calibration, and semantic validation therefore fulfill distinct functions [7]. The proposed provenance record supports the concurrent application of these checks without conflating one as a replacement for the others.

The present study has limited empirical causal scope. Its controlled manipulations characterize the behavior of fixed detector functions applied to archived responses. Attribution at the generation level requires additional paired generations, stable

annotation references, and recorded decoding conditions. Internal causal tracing would further necessitate access to intermediate model states. Extending the framework to these settings would enable unified linkage of evidence across retrieval, generation, and evaluation via consistent identifiers [23]. The current findings support this experimental design while leaving those implementation mechanisms unresolved. Human annotation uncertainty, multilingual text, near-duplicate sources, and transfer to more recent models represent additional limitations. Addressing them would require newly collected validation data and precisely defined evaluation populations rather than extrapolation from the released corpus.

6. Conclusion

This paper presented a framework connecting quantitative hallucination assessment with causal provenance analysis. It separates source-grounded annotation, fitted detection, and modifications to the generation process. An evaluation used 17,790 RAGTruth responses linked to 2,965 sources, retaining the official source-disjoint partition and validating every annotated character interval against its response text.

The corpus analysis found 7,664 responses with marked hallucinations and 14,289 annotated spans. Excluding implicitly true spans reduced the positive response count to 7,010. Response prevalence and character-weighted burden produced different task orderings, demonstrating the need to state the aggregation unit and annotation policy. On the held-out set, combined lexical and metadata logistic regression achieved AUROC 0.873 and average precision 0.801. Metadata alone reached AUROC 0.842, indicating substantial predictive structure outside direct evidence comparison.

Controlled evidence transformations exposed additional detector behavior. Task-matched replacement increased the lexical random forest score by 0.294 on average. Empty evidence decreased the lexical logistic regression score by 0.416, revealing a failure at the boundary of its input distribution. These results concern fixed detectors processing fixed responses. They do not estimate how alternative evidence would change newly generated answers or establish which internal language-model component produced an unsupported claim.

The framework provides a reproducible path from archived artifacts to measurements, sensitivity checks, and explicitly scoped modification records. Its empirical limitations include one corpus, lightweight surface features, a fixed donor mapping, and unavailable generation counterfactuals. Further work should evaluate semantic detectors under the same source-aware protocol, validate missing-evidence handling, and collect controlled paired generations with consistent annotation criteria. Such experiments could extend provenance from detector sensitivity to generation mechanisms while preserving the distinction between observed association and an identified causal effect. Independent replication across languages and domains would also clarify which diagnostic patterns persist beyond the task mixture and historical model configurations represented in this public release.

References

1. L. Huang et al., "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions," *ACM Transactions on Information Systems*, vol. 43, no. 2, pp. 1–55, 2025.
2. Z. Ji et al., "Survey of hallucination in natural language generation," *ACM Computing Surveys*, vol. 55, no. 12, pp. 1–38, 2023.
3. J. Maynez, S. Narayan, B. Bohnet, and R. McDonald, "On faithfulness and factuality in abstractive summarization," in **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, July 2020, pp. 1906–1919.
4. J. Li, X. Cheng, X. Zhao, J. Y. Nie, and J. R. Wen, "Halueval: A large-scale hallucination evaluation benchmark for large language models," in **Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing**, Dec. 2023, pp. 6449–6464.
5. S. Lin, J. Hilton, and O. Evans, "TruthfulQA: Measuring how models mimic human falsehoods," in **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, May 2022, pp. 3214–3252.

6. C. Niu et al., "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)*, Aug. 2024, pp. 10862–10878.
7. S. Min et al., "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*, Dec. 2023, pp. 12076–12100.
8. P. Manakul, A. Liusie, and M. Gales, "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*, Dec. 2023, pp. 9004–9017.
9. J. Pearl, *Causal Inference in Statistics: An Overview*, 2009.
10. S. Farquhar, J. Kossen, L. Kuhn, and Y. Gal, "Detecting hallucinations in large language models using semantic entropy," *Nature*, vol. 630, no. 8017, pp. 625–630, 2024.
11. P. Lewis et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 9459–9474.
12. A. Mallen et al., "When not to trust language models: Investigating effectiveness of parametric and non-parametric memories," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, July 2023, pp. 9802–9822.
13. N. F. Liu et al., "Lost in the middle: How language models use long contexts," *Transactions of the Association for Computational Linguistics*, vol. 12, pp. 157–173, 2024.
14. T. Gao, H. Yen, J. Yu, and D. Chen, "Enabling large language models to generate text with citations," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Dec. 2023, pp. 6465–6488.
15. A. Wang, K. Cho, and M. Lewis, "Asking and answering questions to evaluate the factual consistency of summaries," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, July 2020, pp. 5008–5020.
16. A. R. Fabbri, C. S. Wu, W. Liu, and C. Xiong, "QAFactEval: Improved QA-based factual consistency evaluation for summarization," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, July 2022, pp. 2587–2601.
17. P. Laban, T. Schnabel, P. N. Bennett, and M. A. Hearst, "SummaC: Re-visiting NLI-based models for inconsistency detection in summarization," *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 163–177, 2022.
18. O. Honovich et al., "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*, July 2022, pp. 3905–3920.
19. Y. Zha, Y. Yang, R. Li, and Z. Hu, "AlignScore: Evaluating factual consistency with a unified alignment function," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, July 2023, pp. 11328–11348.
20. K. Meng, D. Bau, A. Andonian, and Y. Belinkov, "Locating and editing factual associations in GPT," in *Advances in Neural Information Processing Systems*, vol. 35, 2022, pp. 17359–17372.
21. L. Tang, P. Laban, and G. Durrett, "Minichack: Efficient fact-checking of LLMs on grounding documents," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Nov. 2024, pp. 8818–8847.
22. B. Schölkopf et al., "Toward causal representation learning," *Proceedings of the IEEE*, vol. 109, no. 5, pp. 612–634, 2021.
23. M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug. 2016, pp. 1135–1144.
24. S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.
25. L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
26. A. Niculescu-Mizil and R. Caruana, "Predicting good probabilities with supervised learning," in *Proceedings of the 22nd International Conference on Machine Learning*, Aug. 2005, pp. 625–632.
27. T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," *PLoS ONE*, vol. 10, no. 3, p. e0118432, 2015.
28. C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *International Conference on Machine Learning*, Jul. 2017, pp. 1321–1330.
29. S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in machine-learning-based science," *Patterns*, vol. 4, no. 9, 2023.
30. B. Efron, "Bootstrap methods: another look at the jackknife," in *Breakthroughs in Statistics: Methodology and Distribution*, New York, NY: Springer New York, 1992, pp. 569–593.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of Publisher and/or the editor(s). Publisher and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.