Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Confidence Auditing in Automated Decision Systems

Md. Anamul Hasan

developer.anamulhasan@gmail.com

national university bangladesh https://orcid.org/0009-0005-9925-0855

Md.Abu Sayeed

East West University https://orcid.org/0009-0003-0698-2607

Research Article

Keywords: Conformal prediction, confidence calibration, uncertainty quantification, algorithmic auditing, automated decision systems, expected calibration error, forecast evaluation, distribution-free inference

Posted Date: October 7th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11274063/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

RESEARCH PAPER · FULL MANUSCRIPT

Confidence Auditing in Automated Decision Systems

Validating Reported Confidence Against Realised Error Rates Using Conformal Prediction

Author : Md. Anamul Hasan

Designation : Data Analyst | Statistics & Data Science Background

Orchid ID : 0009-0005-9925-0855

Co-author : Md.Abu sayeed | Founder of Cuteadmoa | AI model and system developer

Orchid ID : 0009-0003-0698-2607

Document type: full research paper · 5 figures · 4 tables · prepared for academic development and supervisor review.

ABSTRACT

BACKGROUND : Automated decision systems routinely attach a confidence figure to their output: a softmax probability, a self-reported certainty, a prediction interval. These figures govern whether an output is trusted, escalated or accepted unreviewed, yet the confidence a system reports and the error rate it actually exhibits are rarely compared statistically.

THE GAP : Calibration research measures and mitigates miscalibration; conformal inference constructs prediction sets with guaranteed coverage; the auditing literature asks for accountability. No established protocol takes a deployed system's own confidence claim, validates it against that system's realised error rate using a distribution-free benchmark, and returns a documented verdict.

METHOD : A multi-domain, multi modality audit protocol covering three confidence modalities and three system classes. Coverage gap, debiased and adaptive calibration error, sharpness and discriminative power are measured with intervals; a split-conformal construction supplies the reference standard; a pre-specified tolerance and exact binomial test define the decision rule for a validated, failed or over-covered claim.

CONTRIBUTION : An inversion of conformal prediction's role used not to improve a model's uncertainty reporting but as an external standard against which a system's own reported confidence is judged and a pre-specified audit procedure that terminates in a per-system verdict reproducible by a third party.

Keywords: Conformal prediction; confidence calibration; uncertainty quantification; algorithmic auditing; automated decision systems; expected calibration error; forecast evaluation; distribution-free inference

1. Introduction

A modern automated decision system rarely issues a bare answer. A classifier returns a label and a probability; a large language model (LLM) answers a question and, when asked, states how sure it is; a forecasting model produces a point estimate wrapped in an interval. The confidence component is the part of the output that tells the human, or the surrounding software, how much weight the answer deserves. It is therefore the component that governs reliance and, increasingly, the component that governs regulatory treatment, since several accountability frameworks treat a system’s ability to signal its own unreliability as central to safe deployment [13], [27].

The difficulty is that confidence is reported far more often than it is checked. A softmax score is an artefact of how a network was trained, not a promise about accuracy; modern deep networks are systematically

overconfident, a result established almost a decade ago by Guo et al. [1] and since replicated across architectures and domains. Self-reported confidence from language models is generated the same way any other token is generated, and has been repeatedly shown to be poorly calibrated, with instruction-tuned models frequently worse than their base counterparts [7], [8], [23]. Forecast intervals are constructed under assumptions that are rarely tested on the realised series they were meant to cover [11], [12].

In other words, the machinery that produces confidence numbers is well developed, and the evidence that those numbers are frequently wrong is well established, but the practice of taking a specific system's confidence claims and validating them against that system's realised errors in a way that yields a defensible yes-or-no answer remains rare. Calibration research typically measures aggregate calibration error on a benchmark and proposes a method to reduce it. Auditing practice, by contrast, needs to interrogate a claim: this system says it is right 90 percent of the time when it reports 90 percent confidence; is that true, and can we demonstrate it?

Conformal prediction is a natural instrument for that interrogation. Developed by Vovk, Gammerman and Shafer [2], [3] and popularised for applied work by Angelopoulos and Bates [4], it converts any predictive model into a set-valued predictor with a finite-sample coverage guarantee that holds without assumptions about the model or the data distribution the single assumption is exchangeability of the calibration and test points. The consequence is important for audit purposes: if a model's confidence claims are honest, conformal coverage computed from its own scores will match the nominal level; if they are not, coverage will fall short by a measurable margin that carries a directly interpretable meaning. Conformal prediction thus supplies something the auditing literature has lacked a distribution-free yardstick against which a confidence claim can be judged, rather than a diagnostic that merely describes miscalibration.

This paper develops that idea into a concrete audit. We address a single organising question: can a system's reported confidence be validated, statistically and reproducibly, against its realised error rate and does conformal prediction provide the right benchmark for doing so? The contribution is deliberately procedural rather than model specific. We specify the confidence modalities to be audited (softmax scores, verbalised LLM confidence, forecast intervals), the system classes to be covered (discriminative classifiers, generative question-answering models, and numerical forecasters), the metrics that constitute a verdict, the conformal constructions that establish the reference standard, and the statistical tests that separate real miscalibration from finite-sample noise. The result is an audit protocol that can be executed against a deployed system and that terminates in a documented finding.

The remainder of the paper is organised as follows. Section 2 reviews the three literatures this work sits between confidence reporting and calibration, conformal prediction, and algorithmic auditing and identifies the gap. Section 3 states the research objectives and questions. Section 4 sets out the proposed methodology in detail, including the audit metrics, the conformal benchmark construction, and the analysis plan. Section 5 describes the expected contributions, Section 6 the significance for practice and policy, Section 7 the limitations and threats to validity, and Section 8 concludes.

1.1 Scope and terminology

We use reported confidence to mean whatever scalar or interval the system itself emits as an indication of certainty, and realised accuracy to mean the empirical correctness of that system's outputs on held out data with known ground truth. Calibration denotes agreement between the two in the aggregate; coverage denotes the proportion of test cases for which the true answer falls inside the region the system implied it would cover.

We use audit in the algorithmic-accountability sense established by Cobbe et al. [13]: a structured, documented examination producing findings that a third party can review. The audit proposed here is a statistical audit of a specific claim, not an evaluation of a model's quality in general.

A note on what this paper is not. It does not propose a new calibration method, and it does not attempt to improve any model. The object of study is the confidence claim itself and the means of testing it. Methods such as temperature scaling [1], conformal calibration [4], and rejection-based abstention [16], [28] appear here as instruments for constructing the benchmark and interpreting deviations from it, not as the contribution.

2. Background and Related Work

Three research traditions bear on this problem, and they have developed largely in parallel. The first studies how automated systems report confidence and how badly those reports deviate from reality. The second supplies distribution free machinery for constructing prediction sets with guaranteed coverage. The third concerns the institutional practice of auditing automated decision systems and the evidence such audits require. We review each in turn, then state the gap.

2.1 How automated systems report confidence

Confidence reaches the user through three distinct channels, and the audit must accommodate all three because their failure modes differ.

Softmax and probabilistic scores. The classical case is a discriminative model that outputs a probability vector. The convention of reading the maximum class probability as confidence is near-universal and, as Guo et al. [1] demonstrated across vision and language tasks, systematically optimistic. Their core finding that depth, width, batch normalisation and weight decay all degrade calibration while improving accuracy reframed miscalibration as a structural property of modern training regimes rather than an occasional oversight. Temperature scaling, a single-parameter post-hoc correction, remains a strong baseline, and the expected calibration error (ECE) they popularised remains the most widely reported diagnostic.

Self-reported confidence from language models. When an LLM is asked to state how confident it is, the resulting number is generated text, not a normalised probability. Xiong et al. [7] evaluated confidence elicitation strategies across several model families and found consistent overconfidence, with verbalised confidence tracking human overconfidence patterns rather than true accuracy. Yang et al. [24] examine verbalised scores directly and show that their usefulness depends heavily on how the score is elicited and evaluated. Wang et al. [6] survey the wider space of confidence estimation and calibration for LLMs and document that the field has many proposed estimators but comparatively few settled results about which are trustworthy in deployment. A recurring and directly relevant observation is that instruction tuning can degrade calibration relative to base pre-training [23].

Prediction intervals and density forecasts. In numerical forecasting, confidence is expressed as an interval or a predictive distribution, and its quality is assessed by coverage and sharpness. The econometric literature has developed the relevant diagnostics for decades: Diebold, Gunther and Tay [11] formalised density forecast evaluation by testing whether the probability integral transform of realised values is uniform, which is precisely a test of whether stated predictive distributions are honest. Casillas Olvera and Diebold [12] applied probability orecast evaluation to central bank forecasts, motivated by exactly the accountability concern that drives this paper. The important structural point is that interval confidence is only as good as the model

assumptions behind homoskedastic residuals, correct functional form, stable parameters and those assumptions are rarely audited on the realised series.

Table 1. The three confidence modalities, how each is produced, and its characteristic failure mode.

Modality	System class	Production mechanism	Characteristic failure
Softmax / class probability	Discriminative classifier	Product of the training objective	Systematic overconfidence; worsens with depth, width, batch normalisation and weight decay
Verbalised self-reported score	Generative question answering	Generated as text, like any other token	Instruction tuning degrades it; highly sensitive to elicitation format and prompt wording
Prediction interval / density	Numerical forecaster	Built on model assumptions about residuals	Under-coverage under shift; heteroskedastic residuals break the stated width

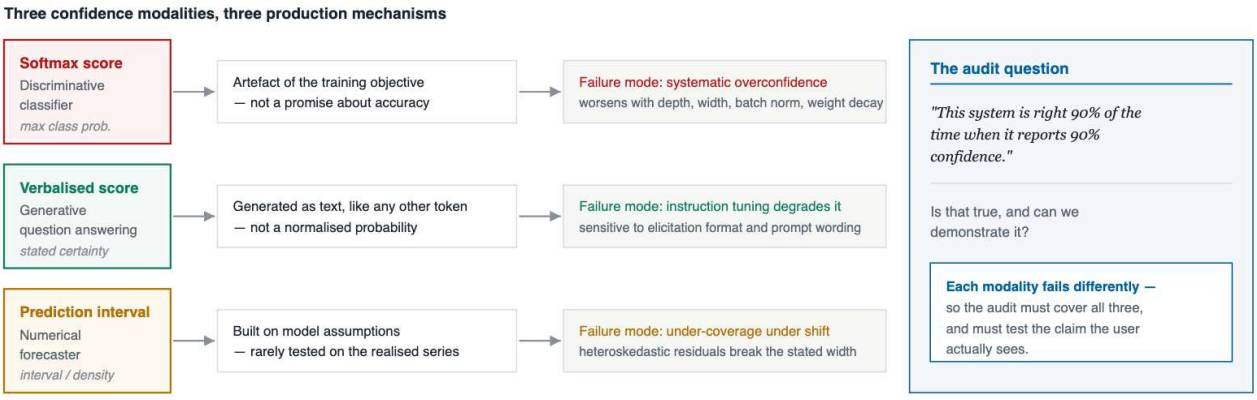


Figure 1. The three confidence modalities, how each is produced, and the failure mode specific to each. The audit covers all three because their failure modes are structurally different, and tests the figure the user actually sees rather than an internal score.

2.2 Measuring calibration, and the limits of the measurement

The standard diagnostic is the expected calibration error: predictions are binned by confidence, and ECE is the average absolute difference between mean confidence and empirical accuracy within each bin [1]. It is intuitive, cheap, and near-universally reported and it has known weaknesses that matter for an audit.

ECE is sensitive to the number of bins and to the choice of binning scheme, and it is biased in a way that does not vanish with sample size. Kumar, Liang and Ma [5] addressed this directly, proposing debiased estimators that correct the binning bias and show that some reported calibration improvements are artefacts of estimator choice rather than genuine gains. This result is central to the present argument: if the measurement instrument is itself unreliable, an audit built on it cannot produce a defensible verdict. It is one reason we treat conformal

coverage, whose guarantee is analytic rather than estimator-dependent, as the reference standard and ECE as a descriptive companion.

A second limitation is that aggregate calibration is not the property a deployment actually depends on. A model can be well calibrated on average across a benchmark while being badly overconfident on the specific subpopulation a given deployment serves. Ovadia et al. [15] showed that calibration degrades substantially under dataset shift, and a body of work has since examined how to recover calibrated uncertainties when the test distribution differs from training [14], [15]. For audit purposes the implication is that aggregate metrics must be supplemented by conditional ones coverage within groups and within confidence strata because the failure that causes harm is usually conditional, not average.

A third limitation is that calibration says nothing about whether confidence is *useful*. A model that reports 0.7 confidence on every case can be trivially calibrated if its accuracy is 0.7, yet its confidence conveys no information and supports no triage decision. Discriminative power the extent to which confidence separates correct from incorrect outputs, typically measured by the area under the ROC curve treating correctness as the label is therefore a necessary complement to calibration. A useful audit reports both, because a system can pass a calibration check while failing as a decision aid.

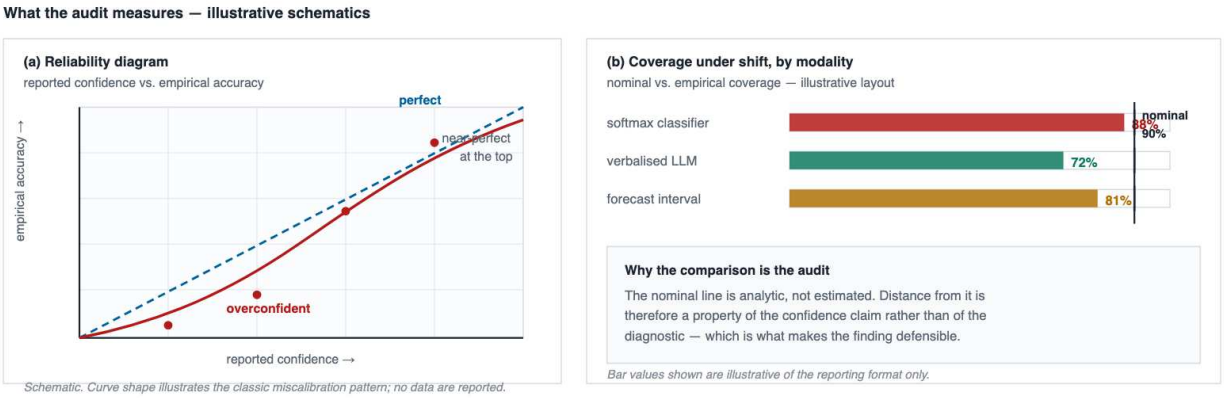


Figure 2. What the audit measures. Panel (a) is the reliability diagram in schematic form; panel (b) shows the coverage comparison at the nominal level, the quantity carrying the analytic guarantee. Both panels illustrate the reporting format only.

2.3 Conformal prediction as a distribution-free benchmark

Conformal prediction originates with Vovk, Gammerman and Shafer [2] and is developed in tutorial form by Shafer and Vovk [3] and, for a machine-learning audience, by Angelopoulos and Bates [4]. The construction is simple. Given a fitted model, a held out calibration set, and a nonconformity score that measures how poorly each calibration point conforms to the model, one computes the empirical quantile of the calibration scores at the level implied by the target coverage and uses it to build a prediction set for each new input. The resulting set contains the true label with probability at least the nominal level, in finite samples, provided only that the calibration and test points are exchangeable. No assumption is made about the model's correctness, its architecture, or the shape of the data distribution.

Three properties make conformal prediction well suited to auditing rather than merely to prediction. First, the coverage guarantee is analytic, so a shortfall is interpretable: if a method claiming 90 percent coverage delivers 74 percent on exchangeable held-out data, the discrepancy is a property of the method's confidence claim, not

of the diagnostic. Second, the framework accommodates arbitrary nonconformity scores, which means the same machinery can audit a softmax classifier, an LLM answering multiple-choice questions, and a regression forecaster, using scores appropriate to each [4], [25]. Third, variants extend the guarantee to settings the base construction cannot handle. Lei et al. [26] established distribution-free predictive inference for regression, Gibbs and Candès [21] extended conformal inference to distribution shift, Romano, Patterson and Candès [30] introduced conformalised quantile regression for heteroskedastic targets, and Bates et al. [22] generalised coverage to risk control. These extensions matter because an audit that only works under exchangeability would be of limited practical use.

Conformal prediction has already been applied to language models. Kumar et al. [9] used it to quantify uncertainty in multi-choice question answering; Campos et al. [10] surveyed its use across natural language processing; and subsequent work has explored conformal methods for validity guarantees on LLM output [29]. Characteristically, this literature uses conformal prediction to improve a model's uncertainty reporting. Our use is the inverse: we use it as an external standard against which a system's own reported confidence is judged. That inversion is the conceptual move this paper makes.

2.4 Auditing automated decision systems

The third tradition concerns accountability. Cobbe, Lee and Singh [13] introduced reviewability as a framework for making automated decision-making accountable, arguing that systems must be interrogable by parties who did not build them which presupposes that the quantities a system reports can be checked against independent evidence. Regulatory instruments have moved in the same direction: the EU AI Act [27] imposes conformity assessment, documentation and human oversight obligations on high-risk systems, all of which depend on measurable claims about system behaviour.

A parallel literature examines what happens when confidence reaches a human. Miscalibrated AI confidence has been shown to impair appropriate reliance and reduce the efficacy of AI-assisted decisions [17], and verbalised uncertainty influences user trust in ways that do not always track accuracy [18]. At the organisational level, the practical consequences of unverified AI output are documented in adjacent domains Sarkar [19] frames generative AI assistance in professional software as a risk to deeply learned expert practice, and the spreadsheet error literature [20], [31] established decades ago that unverified computational output carries material error rates even when produced entirely by humans. The audit proposed here is motivated by the same concern: a confidence figure that has never been validated against realised error is an unverified claim circulating inside a decision process.

2.5 The gap

The literature establishes three things firmly. Confidence reporting is ubiquitous and takes three structurally different forms. Miscalibration is pervasive and well documented, particularly for deep networks and language models. Conformal prediction provides distribution free finite-sample coverage guarantees that are analytically interpretable.

What is missing is the connection. Calibration research measures and mitigates; conformal research constructs guaranteed sets; auditing research asks for accountability but no established protocol takes a deployed system's own reported confidence, validates it against that system's realised error rate using a distribution-free benchmark, and returns a documented verdict covering both calibration and coverage. Two consequences follow from the gap. Practitioners have no standard instrument for answering the question procurement and

oversight actually pose, namely whether a stated confidence level means what it says. And the audit evidence that accountability frameworks imply should exist a testable claim about reliability, with a result is not being produced.

This paper addresses that gap by specifying the audit. The following section states the objectives and research questions the audit is designed to answer.

3. Research Objectives and Questions

The overarching objective is to establish whether a deployed automated decision system's reported confidence can be validated against its realised error rate using a distribution free benchmark, and to specify the audit that performs that validation. Four research questions operationalise the objective.

RQ1 — Magnitude of the discrepancy. How large is the gap between reported confidence and realised accuracy, measured both by conventional calibration metrics and by conformal coverage shortfall, across the three confidence modalities and the three system classes under audit? The question is descriptive in form but consequential in effect: a measured shortfall is the audit's primary finding.

RQ2 — Modality and system dependence. Does the discrepancy vary systematically by confidence modality (softmax probability, verbalised LLM confidence, forecast interval), by system class (discriminative classifiers, generative question answering, numerical forecasting), and by domain? If verbalised confidence is systematically worse than softmax confidence, or if interval coverage degrades faster under distribution shift than probabilistic coverage, those are actionable results for how systems should be instrumented.

RQ3 — Does the benchmark close the gap, and at what cost? When the conformal construction is applied to a system that fails the audit, does coverage recover to the nominal level, and what is the price in sharpness that is, how much larger and less informative do the resulting prediction sets become? A remedy that achieves nominal coverage by producing sets too broad to inform any decision has not solved the deployment problem. The question is therefore inherently two-sided: validity and efficiency must be reported together.

RQ4 — Is reported confidence informative for triage? Setting calibration aside, does reported confidence discriminate between correct and incorrect outputs well enough to support an escalation rule for example, routing cases below a confidence threshold to human review? This is measured by the area under the ROC curve treating correctness as the label, and by the trade-off between the fraction of cases escalated and the fraction of errors caught. A system may be miscalibrated yet still useful for triage, or calibrated yet uninformative; the audit must distinguish these cases rather than collapse them into a single verdict.

The four questions are ordered so that they compose into a single audit report. RQ1 establishes whether a discrepancy exists; RQ2 characterises where it concentrates; RQ3 tests whether a documented remedy restores the claim and at what cost; RQ4 asks whether the confidence signal retains operational value even when the claim fails.

Two objectives are deliberately excluded. The audit does not attempt to explain why a given model is miscalibrated in mechanistic terms, and it does not rank models by quality. Its purpose is to test a claim, not to improve a model or to select among competing models.

Table 2. Research questions mapped to primary metrics and decision outputs.

RQ	Question in brief	Primary metrics	Decision output

RQ1	How large is the gap between reported confidence and realised accuracy?	Coverage gap with bootstrap interval; debiased expected calibration error	Pass or fail against the pre-specified tolerance
RQ2	Does the gap vary by modality, system class and domain?	All metrics stratified by modality, class and domain	Where miscalibration concentrates
RQ3	Does the conformal benchmark close the gap, and at what cost?	Corrected coverage; mean prediction-set size	The remedy priced in sharpness
RQ4	Does reported confidence still support triage?	AUROC of confidence for correctness; escalation-versus-error trade-off	Whether the signal retains operational value

3.1 Hypotheses

Each research question is expressed as a directional hypothesis below, fixed before the audit is executed. The primary hypothesis is H1, because a measured coverage shortfall is the audit's finding; H2 to H4 determine how that finding is characterized and what can be done about it.

H1 (magnitude — primary). Reported confidence exceeds realised accuracy across all three modalities, and the conformal coverage gap is materially larger than finite sample variation would explain at the audited level. This is the audit's central claim: that the confidence systems report is not, in general, a valid statement about their own error rates.

H2 (modality dependence). The shortfall is not uniform across modalities. Verbalised confidence from language models is expected to depart furthest from nominal, softmax scores next, and forecast intervals least when the target series is stationary on the reasoning that the further a confidence figure sits from a calibrated probabilistic construction, the more it depends on model behaviour that calibration research has repeatedly shown to be optimistic.

H3 (remedy and cost). Applying the conformal construction restores empirical coverage to within tolerance of nominal, at a measurable cost in sharpness. The hypothesis is directional on both parts: coverage recovers, and prediction sets or intervals widen. A remedy that recovered coverage without any loss in sharpness would be surprising and would warrant scrutiny of the construction.

H4 (triage value). Reported confidence retains discriminative power for correctness even where it fails as a calibrated probability, such that a confidence threshold can support an escalation rule better than chance. If supported, a system with a failed confidence claim may still be operationally useful a distinction that matters, because conflating the two would discard workable systems.

The hypotheses are separable by construction. H1 can hold while H4 also holds; a system may be badly miscalibrated and still order its own errors usefully. The audit reports both rather than resolving them into a single verdict, and the report indicates for each hypothesis whether it was supported.

4. Proposed Methodology

4.1 Design

The audit is a multi-domain, multi-modality validation study with a fixed protocol. Each system under audit is evaluated three times on the same held out data: once against its own reported confidence using conventional calibration metrics, once against conformal coverage computed from its own scores, and once against a conformal benchmark constructed independently of the system's reported confidence. The first two measure the system's claim; the third establishes the reference against which the claim is judged. The design is comparative across modalities and system classes, and the unit of analysis is the individual prediction.

The protocol is specified before execution, with the audit level, tolerance, metrics, and decision rule fixed in advance. This pre-specification matters because an audit that selects its metrics after seeing results is not an audit. We therefore treat the protocol itself as a deliverable of the work, alongside the empirical findings.

4.2 Systems under audit

Three system classes are covered, chosen because they represent distinct confidence mechanisms rather than because they are convenient.

Class A — Discriminative classifiers. Supervised models producing class probabilities, including at least one deep network and one classical model such as gradient-boosted trees or logistic regression. The classical models matter as controls: the calibration literature predicts they will be better calibrated than the deep networks [1], and an audit that cannot reproduce that known contrast is not measuring what it claims.

Class B — Generative question-answering models. Language models answering questions with verifiable ground truth, in at least two elicitation conditions: a continuous verbalised confidence score, and a binary or Likert certainty statement. Elicitation format is treated as an experimental factor because prior work shows it materially affects the quality of the reported confidence [7], [24].

Class C — Numerical forecasters. Regression and time-series models producing prediction intervals, evaluated on realised series with known outcomes. This class is where coverage is the conventional criterion and where the econometric evaluation literature supplies established tests [11], [12].

4.3 Task suites and ground truth

Each system class is audited on at least two held-out task suites with unambiguous ground truth, and each suite is partitioned into a calibration split and a test split by random assignment. Where the suite permits, we additionally construct a shifted test split a domain, time period, or subpopulation deliberately not represented in the calibration split to test whether the audit verdict survives distribution shift, following Ovadia et al. [15] and Gibbs and Candès [21].

Ground truth must be objective and independently established. Tasks whose correctness depends on subjective judgement are excluded, because an audit whose ground truth is itself contestable cannot produce a defensible verdict. Where feasible, a second annotator resolves a random sample of cases to quantify labelling error, which propagates into the audit's confidence bounds.

4.4 The conformal benchmark

The benchmark is a split-conformal construction applied to the system's own outputs, chosen so that the guarantee is analytic rather than asymptotic.

Step 1 — Score construction. A nonconformity score is defined per system class. For classifiers, the score is conventionally one minus the probability assigned to the true class. For generative question answering with

multiple choice structure, the score is one minus the model's probability for the true option where logits are available, and a rank-based score over candidate answers where they are not. For forecasting, the score is the absolute residual from the point prediction. The score definition is fixed before execution and reported in full.

Step 2 — Calibration quantile. The empirical quantile of calibration-set scores is computed at the level implied by the target coverage, using the finite-sample correction that accounts for the size of the calibration set. This is the only quantity drawn from data; the guarantee follows from the exchangeability of calibration and test points and requires no assumption about the model.

Step 3 — Coverage measurement. Prediction sets are constructed on the test split and empirical coverage is recorded. The coverage gap nominal coverage minus empirical coverage is the audit's primary quantitative finding. A small positive gap is expected from finite sample variation; the audit's decision rule is calibrated to that variation, as specified in Section 4.6.

Step 4 — Benchmark against the system's own claim. Two comparisons are made. Comparing conformal coverage to nominal level tests whether the model's underlying scores support the coverage its confidence implies. Comparing the system's reported confidence directly against realised accuracy tests the claim the user actually sees. A system can fail the second while passing the first, for example when a model is internally well behaved but reports a transformed or rounded confidence figure and that pattern is itself an audit finding, because the number reaching the user is the number that governs reliance.

Extensions are applied where the base construction's exchangeability assumption is untenable. Group-conditional conformal is used to test coverage within subpopulations, conformalised quantile regression [30] where forecast residuals are heteroskedastic, and an adaptive conformal variant [21] where the test stream is non-exchangeable.

Table 3. Conformal constructions used and the failure each guards against.

Construction	Applied when	What it protects against
Split conformal (base)	Calibration and test points exchangeable	Makes no assumption about the model; baseline coverage guarantee
Group-conditional conformal	Subpopulation coverage matters	Aggregate coverage masking a failing subgroup
Conformalised quantile regression	Forecast residuals heteroskedastic	Constant-width intervals that misstate uncertainty
Adaptive conformal inference	Test stream not exchangeable	Silent coverage loss under distribution drift

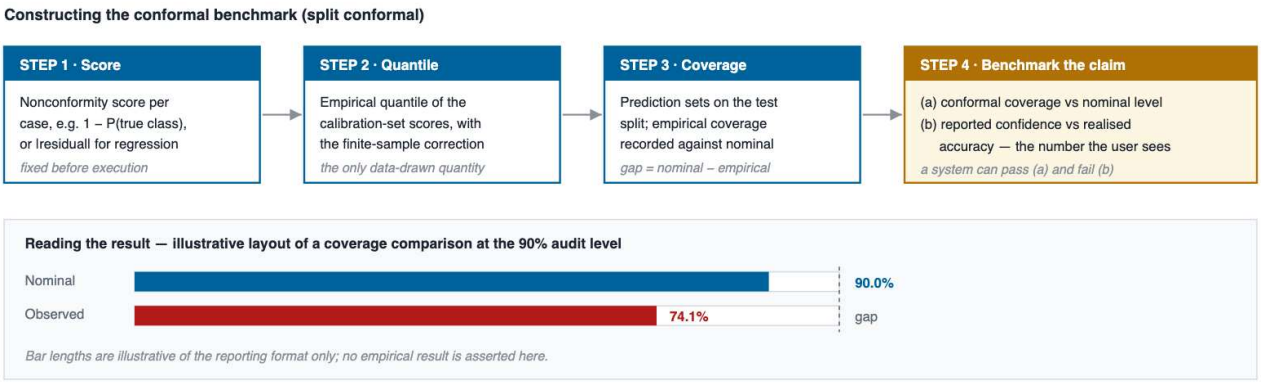


Figure 3. Constructing the conformal benchmark. Step 4 makes the comparison two-sided: a system can satisfy the conformal coverage check (a) while still failing the claim the user sees (b).

4.5 Audit metrics

Five metrics constitute the verdict. They are reported jointly; no single metric is treated as sufficient.

Table 4. Audit metrics, what each measures, and its role in the verdict.

Metric	What it measures	Role in the audit
Coverage gap	Nominal coverage minus empirical conformal coverage, with a bootstrap confidence interval	Primary finding. A gap exceeding the pre-specified tolerance is a failed confidence claim.
Expected calibration error (debiased)	Average absolute difference between reported confidence and empirical accuracy across confidence bins	Descriptive. Reported using the debiased estimator to avoid the binning bias documented by Kumar et al. [5].
Adaptive calibration error	Calibration error restricted to the confidence stratum a deployment actually operates in	Conditional finding. Detects miscalibration that aggregate metrics conceal.
Sharpness / efficiency	Mean prediction-set size, or mean interval width, at the audited coverage level	Cost of the remedy. Nominal coverage achieved only by uninformative sets is reported as such.
Discriminative power	Area under the ROC curve treating correctness as the label; escalation-error trade-off curve	Triage finding. Determines whether confidence retains operational value even where the claim fails.

4.6 Statistical analysis and the decision rule

The audit must separate genuine failure from finite-sample noise, and it must do so before results are seen. Three elements constitute the analysis plan.

Pre-specified tolerance. The maximum coverage gap treated as consistent with a valid claim is fixed in advance, informed by the binomial variability expected at the chosen sample size and confidence level. With a calibration set of sufficient size, the sampling standard error of empirical coverage is small enough that a tolerance can be set tight enough to be meaningful, and sample sizes are chosen accordingly. Systems are

audited on datasets large enough that the tolerance is materially below the discrepancies the literature would predict.

4.6.1 Statistical precision and audit power

The audit's decision rule requires a tolerance wide enough to accommodate sampling noise and narrow enough to be meaningful, and these two requirements trade off directly against the size of the held out test set. The design therefore specifies the precision each choice of sample size delivers, so that the achieved resolution is reported alongside the verdict rather than assumed.

For a coverage audit at the 90% level, the standard error of empirical coverage is $\sqrt{(0.9 \times 0.1 / n)}$. At $n = 500$ test cases this is 1.34 percentage points, giving a 95% interval of roughly ±2.6 points; at $n = 2{,}000$ it falls to 0.67 points, with an interval of roughly ±1.3 points. A tolerance of three percentage points is therefore meaningful at $n = 500$ and comfortably tight at $n = 2{,}000$, while an audit conducted at $n = 150$ would carry a standard error near 2.4 points and could not distinguish genuine miscalibration from noise without a tolerance so wide as to be uninformative.

This yields a concrete planning rule that the audit states in advance: a minimum of 500 held-out cases per system–suite combination, with 2,000 preferred where the tolerance is intended to be tight. Coverage is additionally assessed with an exact binomial test rather than a normal approximation, since at the scale of a single audit the asymptotic approximation is not guaranteed to hold.

The companion consideration is audit power under a null. A non rejection at small sample size is weak evidence of validity, not a finding of it, and the report therefore records the achieved precision for every verdict. A system audited at $n = 200$ and found within tolerance is described as having insufficient evidence to establish failure rather than as validated a distinction that matters because the practical consequence of the two readings is entirely different.

Uncertainty quantification. All metrics are reported with bootstrap confidence intervals obtained by resampling test cases. Coverage is additionally tested against nominal level using an exact binomial test, which requires no asymptotic approximation and is therefore appropriate at the scale of a single audit.

Comparisons across conditions. Differences between modalities, system classes and domains are tested using mixed-effects models with random effects for task suite and system instance, so that the modality effect is not confounded with the particular tasks a modality happens to be audited on. Multiplicity across the family of comparisons is controlled using a false discovery rate procedure, and both raw and adjusted results are reported.

Decision rule. A system’s confidence claim is recorded as validated if the coverage gap lies within the pre-specified tolerance and the exact binomial test does not reject nominal coverage at the corrected significance level. It is recorded as failed if the gap exceeds tolerance in the direction of under-coverage. A claim that exceeds nominal coverage materially over-coverage is recorded separately rather than as a pass, because systematically conservative confidence is a distinct defect: it inflates prediction sets and degrades the utility of the system even though it does not endanger the user. Every verdict is accompanied by the sharpness and discriminative-power figures, so that a failed claim is never reported without the context needed to act on it.

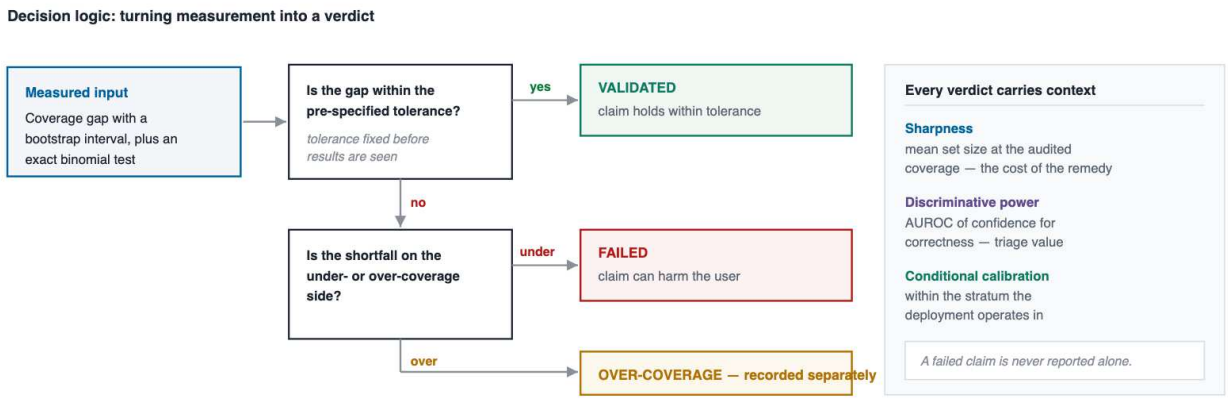


Figure 4. Decision logic. The gate is the pre-specified tolerance; the branch distinguishes under-coverage, which can harm the user, from over-coverage, recorded separately rather than as a pass.

4.7 Procedure, documentation and reproducibility

The audit proceeds in six stages, and the output of each is recorded so that a third party can review the chain of reasoning rather than only its conclusion.

Stage 1 — Claim capture. The system's confidence claim is written down in falsifiable form: what the system asserts about its reliability, on which population, at which confidence level. A claim that cannot be stated in testable terms is itself a finding, since it cannot be audited.

Stage 2 — Data freeze. Task suites, splits, ground truth and the pre specified tolerance are frozen and versioned. Any subsequent change is recorded as an amendment with a stated reason.

Stage 3 — Measurement. The five metrics are computed for every system–suite–split combination, with intervals.

Stage 4 — Remedy test. Where a claim fails, the conformal construction is applied and coverage and sharpness are re-measured, so the audit reports what correction would cost as well as what it would fix.

Stage 5 — Verdict assembly. Findings are assembled into a per-system report containing the claim, the measurements, the decision, and the operational implications.

Stage 6 — Reproduction. All code, score definitions, splits and random seeds are released so that the audit can be re-run. Reproducibility is treated as a requirement rather than a courtesy: an audit finding that cannot be independently reproduced has no standing in an accountability process.

Two threats to validity are addressed procedurally. To prevent the audit from becoming a benchmark leaderboard, results are reported per system against its own claim rather than as comparative rankings. To prevent ground-truth contamination, the audited systems' training data are documented where available, and task suites with plausible training overlap are flagged.

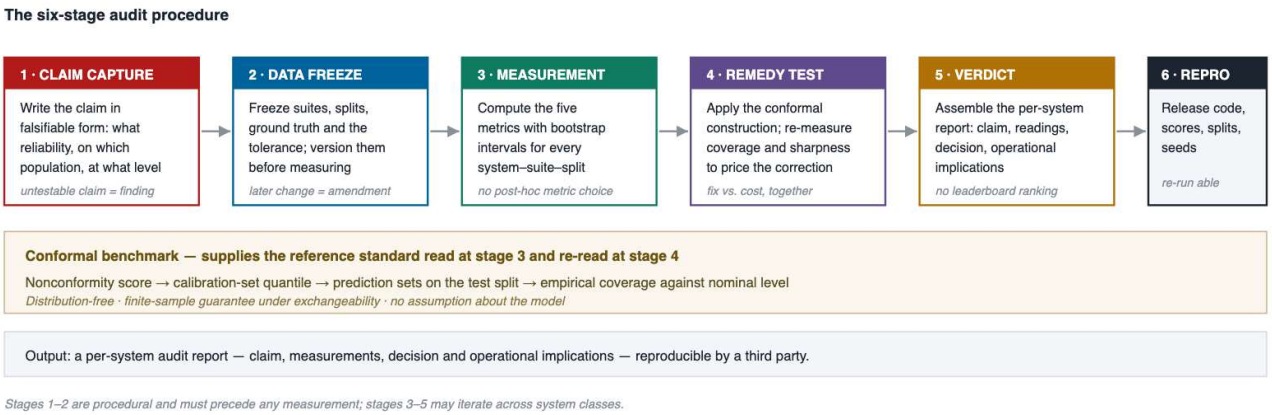


Figure 5. The six-stage audit procedure and the conformal benchmark supplying the reference standard. Stages 1 and 2 must precede any measurement; stages 3 to 5 may iterate across system classes.

5. Expected Contributions

The work is expected to contribute on four fronts.

An audit protocol, specified before execution. The primary contribution is the protocol itself: a pre specified, reproducible procedure that takes a system's stated confidence claim and returns a documented verdict, covering calibration, conditional calibration, coverage, sharpness and discriminative power. Existing practice offers diagnostic metrics and corrective methods; it does not offer a procedure that terminates in a finding about a specific claim, and the absence of such a procedure is a practical obstacle to the accountability frameworks described in Section 2.4.

An inversion of conformal prediction's role. Conformal methods are customarily applied to improve a model's uncertainty reporting. Using conformal coverage as an external benchmark against which a system's own reported confidence is judged is a different use of the same machinery, and it is a use for which the framework is unusually well matched, since its guarantee is analytic and therefore interpretable as a statement about the confidence claim rather than about the diagnostic.

Evidence on where miscalibration concentrates. By auditing three modalities and three system classes on matched suites, the study can report whether discrepancies concentrate in a particular reporting channel for instance whether verbalised LLM confidence is systematically worse than softmax confidence, and whether interval coverage degrades faster under shift than probabilistic coverage. Prior work establishes that miscalibration is widespread; it does not establish a comparative picture across reporting channels using a common benchmark.

A separation of calibration from usefulness. Reporting calibration and discriminative power jointly, and treating over-coverage as a defect rather than a pass, produces a more decision-relevant picture than a single calibration figure. A system may be miscalibrated yet useful for triage, or calibrated yet uninformative, and treating these as the same outcome misdirects both procurement and remediation.

6. Significance

The practical significance follows from a simple observation about how confidence is consumed. Confidence numbers are used to set thresholds, to define escalation rules, and to justify accepting an output without review.

Each of these uses treats the number as a probability statement about correctness. When that statement has never been tested against the system's realised errors, an unverified quantity is silently propagating through a decision process, and the resulting exposure is difficult even to detect because the system continues to behave plausibly.

For organisations deploying automated decision systems, the audit supplies an instrument they currently lack: a way to answer the question procurement actually poses whether a stated confidence level means what it says with a defensible result rather than a vendor assurance. Because the protocol reports the cost of remediation alongside the finding, it also supports a more honest deployment decision than a binary pass or fail: an organisation can see that a system's claims fail, that conformal correction restores coverage, and that the correction broadens prediction sets to the point where human review is required in a stated fraction of cases.

For oversight, the significance is evidentiary. Accountability frameworks presuppose that claims about system behaviour can be verified [13], [27]. A protocol that tests a confidence claim against realised error, with pre-specified tolerances and reproducible code, produces exactly the kind of artefact such frameworks require, and does so without demanding access to model internals which matters, since audits are frequently conducted by parties who cannot inspect the model they are auditing.

For research, the contribution is a bridge. Calibration research, conformal inference and algorithmic auditing have developed separately; the audit protocol connects them by making a calibration claim the object of a conformal test conducted to auditing standards. That connection also opens a line of work on audit-level questions rather than method-level ones: how tight a tolerance is achievable at a given sample size, how audit verdicts behave under distribution shift, and what the minimum evidence is for a confidence claim to be considered validated.

7. Limitations and Threats to Validity

Four limitations should be stated at the outset, because each bears on how audit findings may legitimately be used.

Exchangeability and its violation. The audit's benchmark rests on exchangeability between calibration and test data. Real deployments frequently violate this assumption: populations drift, tasks shift, and the data a system encounters in production is rarely drawn from the same distribution as the data used to construct its calibration set. Conformal variants for distribution shift [21] and heteroskedastic targets [30] mitigate the problem, but mitigation is not elimination: an audit verdict obtained under approximate exchangeability is a statement about the conditions tested, not a permanent certification. Audit findings must therefore be dated and re-established when the deployment context changes.

Ground-truth dependence. Every metric in the audit depends on ground truth. Where labelling is difficult, ambiguous or contested, the audit inherits that ambiguity at its foundation; a coverage gap computed against unreliable labels is not a trustworthy finding. The mitigation is to restrict the audit to tasks with objective, independently established ground truth and to quantify residual labelling error on a sample, propagating it into the reported intervals. This necessarily narrows the domain of tasks to which the audit applies.

Finite-sample tolerance and audit power. The decision rule requires a tolerance wide enough to accommodate sampling noise and narrow enough to be meaningful. At small sample sizes these requirements conflict, and an audit conducted with insufficient data can only detect gross miscalibration. Reported verdicts

must therefore be accompanied by the achieved precision, and a non-rejection at low sample size must be described as insufficient evidence rather than as validation.

Time-bounded conclusions. Systems change. A confidence claim audited against a specific model version, prompt template, or data vintage does not transfer to a subsequent version, and the pace of change in language-model deployment makes this a live concern for the Class B systems in particular. Audit findings should record the exact version, configuration and date tested, and any change to those should be treated as invalidating the prior verdict until re-tested.

Two further threats concern the study design rather than the method. Confidence elicitation for language models is sensitive to prompt wording, and although elicitation format is treated as an experimental factor, unmeasured prompt effects may remain. And the controlled conditions under which audit data are gathered held out suites with known ground truth differ from production conditions, where ground truth often arrives late or not at all; this is the central reason the audit is proposed as a structured exercise rather than as continuous monitoring, and identifying how it might be adapted to streaming production settings remains future work.

8. Conclusion

Automated decision systems attach confidence figures to their outputs, and those figures govern whether the outputs are trusted, escalated or accepted unreviewed. The systems that produce them are known to be miscalibrated, yet the confidence a given system reports is rarely tested against the errors that system actually makes. This paper proposes to close that gap with a structured audit that treats the confidence claim as the object of study and adopts conformal prediction as the benchmark against which the claim is judged.

The choice of conformal prediction is not incidental. Its coverage guarantee is distribution-free and finite-sample, which means a shortfall between nominal and empirical coverage is interpretable as a failure of the confidence claim itself rather than as an artefact of the diagnostic. That interpretability is precisely what an audit needs and what existing calibration metrics, useful as they are descriptively, cannot supply on their own.

The audit specified here covers three confidence modalities and three system classes, fixes its tolerance and decision rule in advance, reports coverage, conditional calibration, sharpness and discriminative power jointly, and terminates in a documented per-system verdict accompanied by the cost of remediation. Its central claim is modest but consequential: before a confidence figure may reasonably be relied upon, it should be tested against realised error, and the test should be one a third party can reproduce. Confidence that has never been audited is not a guarantee. It is a claim, and claims about reliability are the kind that should be checked.

Author note. Correspondence concerning this manuscript should be addressed to the corresponding author. The audit protocol described here is a proposed methodology; no empirical results are asserted.

References

[1] Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. *Proceedings of the 34th International Conference on Machine Learning (ICML)*, PMLR 70, 1321–1330. arXiv:1706.04599

[2] Vovk, V., Gammerman, A., & Shafer, G. (2005). *Algorithmic Learning in a Random World*. Springer. ISBN 978-0-387-00152-4.

[3] Shafer, G., & Vovk, V. (2008). A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9, 371–421. arXiv:0706.3188

[4] Angelopoulos, A. N., & Bates, S. (2023). Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4), 494–591. arXiv:2107.07511

[5] Kumar, A., Liang, P. S., & Ma, T. (2019). Verified uncertainty calibration. *Advances in Neural Information Processing Systems (NeurIPS)*, 32.

[6] Wang, et al. (2024). A survey of confidence estimation and calibration in large language models. arXiv:2311.08298. (NAACL 2024.)

[7] Xiong, M., et al. (2024). Can LLMs express their uncertainty? An empirical evaluation of confidence elicitation in LLMs. *International Conference on Learning Representations (ICLR)*, 2024.

[8] Yang, et al. (2025). Mind the confidence gap: Overconfidence, calibration, and distractor effects in large language models. arXiv:2502.11028.

[9] Kumar, B., et al. (2023). Conformal prediction with large language models for multi-choice question answering. arXiv:2305.18404.

[10] Campos, M., et al. (2024). Conformal prediction for natural language processing. *Transactions of the Association for Computational Linguistics (TACL)*, 12.

[11] Diebold, F. X., Gunther, T. A., & Tay, A. S. (1998). Evaluating density forecasts with applications to financial risk management. *International Economic Review*, 39(4), 863–883.

[12] Casillas-Olvera, G., & Diebold, F. X. (2006). Probability forecasting and central bank accountability. *Journal of Policy Modeling*, 28(2), 223–234.

[13] Cobbe, J., Lee, M. S. A., & Singh, J. (2021). Reviewable automated decision-making: A framework for accountable algorithmic systems. *Proceedings of the ACM Conference on Fairness, Accountability and Transparency (FAccT)*, 598–609.

[14] Park, S., et al. (2020). Calibrated prediction with covariate shift via unsupervised domain adaptation. *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, PMLR 108.

[15] Ovadia, Y., et al. (2019). Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. *Advances in Neural Information Processing Systems (NeurIPS)*, 32.

[16] Franc, V., et al. (2023). Optimal strategies for reject option classifiers. *Journal of Machine Learning Research*, 24, 1–49.

[17] Understanding the effects of miscalibrated AI confidence on user trust, reliance and decision-making efficacy. (2024). arXiv:2402.07632.

[18] Xu, Z., et al. (2025). Understanding how LLMs' verbalized uncertainty influences user trust, satisfaction and task performance. *International Journal of Human-Computer Studies*.

[19] Sarkar, A. (2024). When Copilot becomes autopilot: Generative AI's critical risk to knowledge work and a critical solution. arXiv:2412.15030.

[20] Panko, R. R., & Aurigemma, S. (2010). Revising the Panko–Halverson taxonomy of spreadsheet errors. *Decision Support Systems*, 49(2), 235–244.

[21] Gibbs, I., & Candès, E. J. (2021). Adaptive conformal inference under distribution shift. *Advances in Neural Information Processing Systems (NeurIPS)*, 34.

[22] 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), 1–34.

[23] Large language models are overconfident in their own answers. (2026). arXiv:2606.03437.

[24] Yang, et al. (2024). On verbalized confidence scores for LLMs. arXiv:2412.14737.

[25] Conformal prediction as a calibration standard for predictive models. (2025). arXiv:2512.17048.

[26] Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., & Wasserman, L. (2018). Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523), 1094–1111.

[27] European Parliament and Council of the European Union. (2024). *Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)*. Official Journal of the European Union.

[28] Casacuberta, S., et al. (2025). Selective omniprediction and fair abstention. *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

[29] Large language model validity via enhanced conformal prediction methods. (2024). *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[30] Romano, Y., Patterson, E., & Candès, E. J. (2019). Conformalized quantile regression. *Advances in Neural Information Processing Systems (NeurIPS)*, 32.

[31] Powell, S. G., Baker, K. R., & Lawson, B. (2009). Errors in operational spreadsheets. *Journal of Organizational and End User Computing*, 21(3), 24–36. (Tuck School of Business, Dartmouth.)

Author: Md. Anamul Hasan · Data Analyst | Statistics & Data Science Background | Linkedin

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Table1confidenceauditingautomated.png
- Table2confidenceauditingautomated.png
- Table3confidenceauditingautomateddecisionmaking.png
- Table4confidenceauditingautomateddecisionmaking.png