

Article

Fixed-Sequence Confidence-Interval Calibration for Finite-Sample Selective Risk Control

Sijin Dong ^{1,*} , Xinyu Hao ² and Hiroyuki Shinnou ¹¹ Department of Computer and Information Sciences, College of Engineering, Ibaraki University, Hitachi 316-8511, Japan² Business School, Hohai University, Nanjing 211100, China

* Correspondence: 24nd305h@vc.ibaraki.ac.jp

Abstract

Selective prediction requires a statistically reliable rule for deciding which model outputs can be returned while controlling the error rate among accepted predictions. In large language model (LLM) applications, uncertainty scores provide useful ranking information, but directly thresholding such scores does not yield finite-sample risk guarantees. We propose confidence-interval calibration (CIC), a general calibration framework that converts an arbitrary uncertainty score into a risk-controlled selective decision rule. Given a held-out calibration sample, CIC estimates the acceptance-conditioned error rate for a pre-specified sequence of candidate thresholds and constructs one-sided upper confidence bounds using either Hoeffding's inequality or the exact Clopper–Pearson procedure. Candidate thresholds are examined sequentially, and CIC returns the most permissive threshold within the consecutively certified prefix. Under independent and identically distributed calibration and deployment samples, we establish a finite-sample, high-probability guarantee that any returned threshold controls the acceptance-conditioned error rate at the prescribed level. In our experiments, calibration samples on the order of 10^3 examples already provide useful certification in favorable regimes, although the required sample size depends on the target risk level, model error rate, and quality of the uncertainty ranking. Experiments on CommonsenseQA and TriviaQA across seven LLMs demonstrate reliable empirical risk control with favorable acceptance rates across a range of target risk levels.

Keywords: confidence intervals; finite-sample inference; large language models; risk control; selective prediction; uncertainty quantification



Academic Editor: Wei Zhu

Received: 15 August 2026

Revised: 27 September 2026

Accepted: 29 September 2026

Published: 2 October 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Large language models (LLMs) have achieved substantial progress in natural language understanding and generation, and are increasingly used in question answering (QA) and decision-support systems [1,2]. Nevertheless, they may still produce incorrect, unsupported, or misleading responses. In reliability-sensitive applications, the relevant question is therefore not only whether a model is accurate on average, but also whether one can identify a subset of predictions on which the model can operate at a prescribed level of reliability.

Uncertainty quantification (UQ) provides a natural basis for such selective prediction [3]. Selective prediction, also known as classification with a reject option, allows a predictor to abstain on uncertain inputs in order to trade coverage for lower risk on the retained predictions [4]. Given an input and a model response, an uncertainty estimator

assigns a score intended to reflect prediction reliability. Low-uncertainty responses may then be accepted, while high-uncertainty responses are withheld. Existing uncertainty estimators for LLMs include entropy-based, sampling-based, and semantic-consistency measures [5–8]. However, these scores are generally useful as ranking statistics rather than calibrated statistical guarantees. In particular, a useful ranking score need not correspond to a calibrated probability of correctness, in the sense that its numerical confidence values match empirical correctness frequencies [9]. A threshold chosen from empirical accuracy alone may therefore underestimate the population error rate among the predictions that are ultimately accepted [10–12].

This motivates a general statistical calibration problem. Let an uncertainty score induce a nested family of accepted subsets through thresholding. For each candidate threshold, the quantity of interest is the error probability conditional on acceptance, while the operational objective is to retain as many predictions as possible subject to a prescribed risk constraint. This differs from ordinary accuracy estimation, which evaluates all predictions regardless of selection. It also differs from classical split conformal prediction, whose primary objective is typically marginal coverage of a prediction set [13,14]. Recent conformal methods for language models have similarly focused on set-valued generation or coverage of admissible outputs [8,15,16]. In selective prediction, by contrast, the system either returns a point prediction or abstains, so the central inferential target is the population error rate conditional on the selection event.

We propose confidence-interval calibration (CIC), a finite-sample calibration framework that converts an arbitrary uncertainty score into a risk-controlled selective decision rule. As illustrated in Figure 1, CIC uses a held-out calibration sample in which each model response is associated with an uncertainty score and a task-specific binary error indicator. For every threshold in a pre-specified ordered candidate set, CIC estimates the acceptance-conditioned error rate and constructs a one-sided upper confidence bound using either a Hoeffding-style bound or an exact Clopper–Pearson bound [17]. The thresholds are examined in sequence, and testing stops at the first threshold that cannot be certified. CIC therefore returns the most permissive threshold within the consecutively certified prefix, or returns no threshold when the requested risk level cannot be supported by the calibration evidence.

The statistical validity of CIC follows from finite-sample confidence bounds combined with fixed-sequence testing. Under independent and identically distributed calibration and deployment samples, we show that any non-null threshold returned by CIC controls the population acceptance-conditioned error rate at the prescribed level with high probability. Importantly, this guarantee does not require the uncertainty score itself to be probabilistically calibrated in the usual sense of matching predicted confidence to observed correctness frequencies [9], nor does it require the population risk to vary monotonically with the threshold. The uncertainty score only needs to define an ordered family of selective policies, while the confidence-interval layer provides the statistical certification.

We study LLM selective answering as a representative application of this general calibration problem. Experiments on the closed-ended CommonsenseQA benchmark and the open-ended TriviaQA benchmark use seven instruction-tuned LLMs and semantic entropy as the uncertainty score. Across multiple target risk levels, the results show that CIC maintains empirical accepted-answer risk near or below the prescribed level while retaining substantial acceptance rates. The comparison between Hoeffding-style and Clopper–Pearson bounds further illustrates the trade-off between statistical conservativeness and selection efficiency in finite calibration samples.

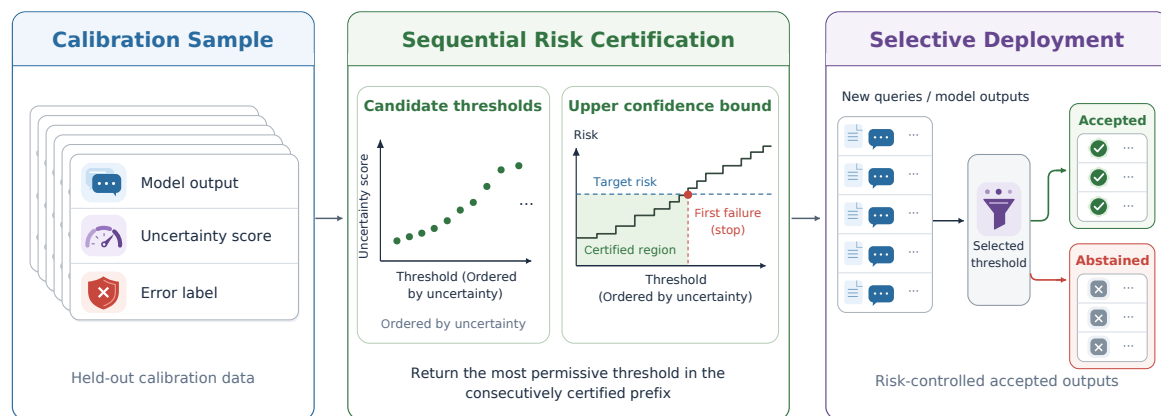


Figure 1. Statistical overview of confidence-interval calibration (CIC). A held-out calibration sample is used to associate each model output with an uncertainty score and a binary error indicator. For a pre-specified sequence of candidate thresholds, CIC estimates the acceptance-conditioned error rate and constructs a one-sided upper confidence bound. Thresholds are examined sequentially, and the most permissive threshold in the consecutively certified prefix is retained. The resulting threshold is then used as a selective decision rule on future samples, providing finite-sample risk control under the stated sampling assumptions. The Hoeffding bound provides a simple distribution-free closed-form certificate, whereas the exact Clopper–Pearson bound exploits the conditional binomial model and is typically less conservative when relatively few selected errors are observed.

The main contributions are summarized as follows:

- We formulate uncertainty-based selective prediction as a finite-sample calibration problem that seeks high acceptance while controlling the population error rate conditional on selection.
- We propose CIC, a general confidence-interval calibration procedure that combines one-sided risk bounds with fixed-sequence threshold testing and can be applied on top of arbitrary uncertainty ranking scores.
- We establish a finite-sample, high-probability guarantee for the data-dependent threshold returned by CIC without requiring monotonicity of the population risk across candidate thresholds.
- We instantiate CIC for selective answering with LLMs and evaluate it on closed-ended and open-ended QA benchmarks across seven models, demonstrating reliable empirical risk control and favorable acceptance rates.

2. Related Work

2.1. Uncertainty Quantification

Uncertainty quantification (UQ) provides numerical measures of prediction reliability and is widely used to rank outputs according to their likelihood of error [3]. For large language models (LLMs), representative approaches include semantic entropy [6], shift attention to relevance [5], and word-sequence entropy [7]. These methods can effectively separate relatively reliable from unreliable responses, making them useful statistics for selective prediction. However, an uncertainty score is generally not calibrated to a population error probability, and a low score does not itself imply a finite-sample guarantee on the error rate of the selected subset. This distinction is related to the broader literature on probabilistic calibration, which studies whether predicted confidence values agree with empirical correctness frequencies [9]. Our framework therefore does not require probabilistically calibrated uncertainty. Instead, it treats the score as an ordering statistic and performs an additional statistical calibration step to certify an operating threshold.

2.2. Conformal Prediction for Language Models

Split conformal prediction uses a held-out calibration sample to construct prediction regions with finite-sample marginal coverage guarantees [13,14]. This principle has been extended to natural language processing [18], including conformal uncertainty for open-ended generation [8,15], conformal language modeling [19], multimodal output selection [16], factuality guarantees [20], and conformal alignment [10]. These methods primarily target coverage of labels, candidate sets, or admissible outputs. The inferential target considered here is different. We study a nested family of point-prediction policies induced by uncertainty thresholds and seek to control the population error probability conditional on acceptance. Thus, the main statistical object is not a prediction set with marginal coverage, but the risk of a data-dependent selective decision rule.

2.3. Risk-Controlled Selective Prediction

Selective prediction allows a model to abstain on uncertain inputs, but reliable deployment requires more than an empirical improvement in accuracy after filtering. Classical work on selective classification studies this risk–coverage trade-off by allowing a predictor to reject uncertain inputs while controlling the error rate on the retained subset [4]. Subsequent methods such as SelectiveNet integrate prediction and rejection into a single learned architecture and directly optimize selective performance [21]. The selected subset should satisfy a user-specified population risk requirement with controlled statistical confidence.

Recent work has studied such risk control for foundation models. COIN calibrates uncertainty thresholds for selective question answering using upper confidence bounds [11], whereas LEC formulates selective prediction and routing through expectation constraints [12]. COIN and CIC both operate as lightweight post hoc calibration layers on top of precomputed model outputs and uncertainty scores, so their calibration cost is typically negligible relative to LLM inference and uncertainty estimation. COIN searches for the largest uncertainty threshold satisfying an upper-confidence-bound criterion, whereas CIC evaluates a pre-specified ordered sequence and stops at the first uncertified candidate. Consequently, CIC requires at most K confidence-bound evaluations for K candidate thresholds, and may terminate earlier under fixed-sequence testing; we therefore view the main distinction from COIN as statistical formulation and threshold-selection structure rather than a substantial difference in computational overhead.

CIC is closely related to this line of work, but emphasizes a direct finite-sample confidence-interval formulation. For a pre-specified ordered sequence of candidate thresholds, CIC constructs one-sided Hoeffding-style or exact Clopper–Pearson upper bounds and applies fixed-sequence certification, stopping at the first uncertified threshold. The resulting rule returns the most permissive threshold in the consecutively certified prefix, yielding high-probability control of the acceptance-conditioned error rate without requiring monotonicity of the population risk across thresholds.

3. Statistical Methodology

3.1. Selective-Risk Formulation

Let $(X, Y^*) \sim \mathbb{P}$ denote an input–ground-truth pair, where $X \in \mathcal{X}$ is an input and $Y^* \in \mathcal{Y}$ is its reference target. A fixed predictive model $\mathcal{L}$ produces $\hat{Y} = \mathcal{L}(X; \xi)$, where ξ denotes possible prediction randomness. Deterministic prediction is included as a special case.

A task-specific evaluation rule $\mathcal{A} : \mathcal{Y} \times \mathcal{Y} \rightarrow \{0, 1\}$ determines whether the prediction is acceptable, with $\mathcal{A}(Y^*, \hat{Y}) = 1$ indicating a correct or acceptable prediction. Define the corresponding binary error indicator as

$$E = 1 - \mathcal{A}(Y^*, \hat{Y}) \in \{0, 1\}. \quad (1)$$

Given a fixed uncertainty measure $\mathcal{M}$, each prediction receives a scalar uncertainty score

$$U = \mathcal{U}(X, \hat{Y}; \mathcal{M}) \in \mathbb{R}, \quad (2)$$

where smaller values indicate greater predicted reliability. For a threshold $t \in \mathbb{R}$, define the selection indicator

$$S_t = \mathbf{1}\{U \leq t\}. \quad (3)$$

A prediction is returned when $S_t = 1$, and otherwise withheld.

The population risk of interest is the error probability conditional on selection,

$$R(t) := \mathbb{P}(E = 1 \mid S_t = 1) = \frac{\mathbb{E}[ES_t]}{\mathbb{E}[S_t]}, \quad (4)$$

whenever $\mathbb{P}(S_t = 1) > 0$. Here, the probability and expectations are taken with respect to the joint randomness induced by $(X, Y^*) \sim \mathbb{P}$ and, when the predictive model is stochastic, the decoding randomness ξ under its fixed sampling distribution. Equivalently, they are taken over the induced population distribution of (U, E, S_t) . We also define the population acceptance probability

$$\pi(t) := \mathbb{P}(S_t = 1) = \mathbb{P}(U \leq t). \quad (5)$$

For a prescribed risk level $\alpha \in (0, 1)$, the statistical objective is to identify a selective rule with large acceptance probability while satisfying

$$R(t) \leq \alpha. \quad (6)$$

Because $\pi(t)$ is non-decreasing in t , larger thresholds correspond to more permissive selective policies. No monotonicity assumption is imposed on the population risk $R(t)$.

We use a calibration sample $\mathcal{D}_{\text{cal}} = \{(X_i, Y_i^*)\}_{i=1}^N$. Applying the fixed predictive model, decoding procedure, uncertainty measure, and evaluation criterion gives $\{(U_i, E_i)\}_{i=1}^N$, where

$$\hat{Y}_i = \mathcal{L}(X_i; \xi_i), \quad U_i = \mathcal{U}(X_i, \hat{Y}_i; \mathcal{M}), \quad E_i = 1 - \mathcal{A}(Y_i^*, \hat{Y}_i).$$

Throughout the paper, we use calibration to refer to the held-out data and the overall threshold-selection procedure, and certification to refer specifically to the decision that a particular candidate threshold satisfies the prescribed risk criterion.

3.2. Finite-Sample Upper Confidence Bounds

For a fixed threshold t , define the number of selected calibration observations as

$$n(t) = \sum_{i=1}^N \mathbf{1}\{U_i \leq t\} \quad (7)$$

and the number of errors among them as

$$m(t) = \sum_{i=1}^N \mathbf{1}\{U_i \leq t\} E_i. \quad (8)$$

The empirical selective risk is

$$\hat{R}(t) = \begin{cases} \frac{m(t)}{n(t)}, & n(t) > 0, \\ 1, & n(t) = 0. \end{cases} \quad (9)$$

Since directly comparing $\hat{R}(t)$ with α ignores finite-sample uncertainty, we instead construct a one-sided upper confidence bound for $R(t)$.

3.2.1. Hoeffding Upper Bound

For $\gamma \in (0, 1)$, define

$$B_H(t; \gamma) = \begin{cases} \min \left\{ 1, \hat{R}(t) + \sqrt{\frac{\log(1/\gamma)}{2n(t)}} \right\}, & n(t) > 0, \\ 1, & n(t) = 0. \end{cases} \quad (10)$$

3.2.2. Clopper–Pearson Upper Bound

The exact one-sided Clopper–Pearson upper confidence bound [17] is

$$B_{CP}(t; \gamma) = \begin{cases} 1, & n(t) = 0 \text{ or } m(t) = n(t), \\ \text{BetaInv}(1 - \gamma; m(t) + 1, n(t) - m(t)), & 0 \leq m(t) < n(t), \end{cases} \quad (11)$$

where $\text{BetaInv}(q; a, b)$ denotes the q -quantile of a $\text{Beta}(a, b)$ distribution. Unlike the distribution-free Hoeffding bound, the Clopper–Pearson construction exploits the conditional binomial model derived below and is often less conservative when few errors are observed.

3.3. Fixed-Sequence Threshold Calibration

Let $\mathcal{T} = \{t_1 < t_2 < \dots < t_K\}$ be an ordered finite set of candidate thresholds. Importantly, $\mathcal{T}$ must be fixed independently of the calibration sample. It may, for example, be specified from a known score range or constructed from an independent pilot sample. This independence is required because the confidence guarantee for each candidate is derived for a threshold that is fixed before observing the calibration data. If candidate thresholds were instead chosen adaptively from the same calibration sample, the selected thresholds would themselves be data-dependent, and the fixed-threshold coverage result could not be applied directly without additionally accounting for this selection.

For each candidate t_k , consider the certification problem

$$H_{0,k} : R(t_k) > \alpha \quad \text{vs.} \quad H_{1,k} : R(t_k) \leq \alpha.$$

A candidate threshold is certified whenever

$$B(t_k; \delta) \leq \alpha, \quad (12)$$

where B denotes either B_H or B_{CP} .

CIC examines the thresholds in increasing order. Testing proceeds only while all previous thresholds have been certified and stops at the first failure. Define

$$\hat{k} = \max \left\{ k \in \{1, \dots, K\} : \begin{array}{l} B(t_k; \delta) \leq \alpha \\ \text{for every } j \leq k \end{array} \right\}, \quad (13)$$

with $\hat{k} = 0$ when the set is empty. The calibrated threshold is

$$\hat{t} = \begin{cases} t_{\hat{k}}, & \hat{k} \geq 1, \\ \text{NULL}, & \hat{k} = 0. \end{cases} \quad (14)$$

Because the acceptance sets are nested,

$$\{U \leq t_1\} \subseteq \{U \leq t_2\} \subseteq \dots \subseteq \{U \leq t_K\},$$

the returned threshold is the most permissive threshold within the consecutively certified prefix. The complete calibration procedure is summarized in Algorithm 1.

Algorithm 1 Confidence-Interval Calibration with Fixed-Sequence Testing

```

1: Input: Calibration pairs  $\{(u_i, e_i)\}_{i=1}^N$ , candidate thresholds  $\mathcal{T} = \{t_1 < \dots < t_K\}$ , risk
   level  $\alpha$ , significance level  $\delta$ 
2: Output: Calibrated threshold  $\hat{t}$ 
3: Initialize  $\hat{t} \leftarrow \text{NULL}$ 
4: for  $k = 1$  to  $K$  do
5:    $n \leftarrow \sum_{i=1}^N \mathbf{1}\{u_i \leq t_k\}$ 
6:    $m \leftarrow \sum_{i=1}^N \mathbf{1}\{u_i \leq t_k\} e_i$ 
7:   Compute  $B(t_k; \delta)$  using (10) or (11)
8:   if  $B(t_k; \delta) \leq \alpha$  then
9:      $\hat{t} \leftarrow t_k$ 
10:  else
11:    break
12:  end if
13: end for
14: Return  $\hat{t}$ 

```

3.4. Finite-Sample Validity

We now establish the statistical validity of the calibration procedure.

Assumption 1 (Independent Calibration Sample). *The calibration pairs $(U_1, E_1), \dots, (U_N, E_N)$ are independently and identically distributed according to a population distribution $P_{U,E}$, and deployment observations follow the same distribution. The predictive model, decoding mechanism, uncertainty estimator, evaluation criterion, and candidate threshold set $\mathcal{T}$ are fixed independently of the calibration sample. If prediction is randomized, the variables $\xi_1, \dots, \xi_N$ are generated independently from the same fixed decoding distribution.*

All probability statements below are with respect to the randomness of the calibration sample. If an independent pilot sample is used to construct $\mathcal{T}$, the statements hold conditionally on that pilot sample.

Theorem 1 (Fixed-Threshold Coverage). *Under Assumption 1, fix any threshold t satisfying $\pi(t) > 0$. Conditional on $n(t) = n > 0$,*

$$m(t) \mid n(t) = n \sim \text{Binomial}(n, R(t)). \quad (15)$$

Consequently, for either $B = B_H$ or $B = B_{CP}$ and every $\gamma \in (0, 1)$,

$$\mathbb{P}_{\text{cal}}(R(t) \leq B(t; \gamma)) \geq 1 - \gamma. \quad (16)$$

Proof. For each calibration observation, define

$$C_i = \begin{cases} 0, & U_i > t, \\ 1, & U_i \leq t, E_i = 0, \\ 2, & U_i \leq t, E_i = 1. \end{cases}$$

Under Assumption 1, the variables $C_1, \dots, C_N$ are independently and identically distributed with

$$\mathbb{P}(C_i = 2) = \pi(t)R(t),$$

$$\mathbb{P}(C_i = 1) = \pi(t)\{1 - R(t)\},$$

and

$$\mathbb{P}(C_i = 0) = 1 - \pi(t).$$

Hence, their category counts follow a multinomial distribution. Conditioning on the total number of selected observations, $n(t) = n$, gives

$$m(t) \mid n(t) = n \sim \text{Binomial}(n, R(t)).$$

For the Clopper–Pearson construction, (16) follows directly from its one-sided binomial coverage property. For the Hoeffding construction, conditional on $n(t) = n > 0$, Hoeffding’s inequality gives

$$\mathbb{P}_{\text{cal}}\left(R(t) > \hat{R}(t) + \sqrt{\frac{\log(1/\gamma)}{2n}} \mid n(t) = n\right) \leq \gamma.$$

Averaging over the possible values of $n(t)$ yields (16). When $n(t) = 0$, the upper bound is defined as one, so the claimed inequality holds trivially. $\square$

Theorem 2 (Fixed-Sequence Risk Control). *Under Assumption 1, let $\mathcal{T} = \{t_1, \dots, t_K\}$ be fixed independently of the calibration sample and let $\hat{t}$ be selected by (13)–(14). Then,*

$$\mathbb{P}_{\text{cal}}(\hat{t} = \text{NULL or } R(\hat{t}) \leq \alpha) \geq 1 - \delta. \quad (17)$$

Proof. If $R(t_k) \leq \alpha$ for every k , the statement is immediate. Otherwise, define the first unsafe candidate as

$$k^* = \min\{k : R(t_k) > \alpha\}. \quad (18)$$

For CIC to return any threshold with population risk greater than α , the fixed-sequence procedure must first certify t_{k^*} . Therefore,

$$\{R(\hat{t}) > \alpha\} \subseteq \{B(t_{k^*}; \delta) \leq \alpha < R(t_{k^*})\} \subseteq \{R(t_{k^*}) > B(t_{k^*}; \delta)\}.$$

Because t_{k^*} is fixed at the population level and does not depend on the calibration sample, Theorem 1 gives

$$\mathbb{P}_{\text{cal}}(R(t_{k^*}) > B(t_{k^*}; \delta)) \leq \delta.$$

Thus,

$$\mathbb{P}_{\text{cal}}(\hat{t} = \text{NULL or } R(\hat{t}) \leq \alpha) \geq 1 - \delta.$$

$\square$

Theorem 2 does not require $R(t)$ to be monotone in the threshold and does not require independence among the confidence bounds computed at different candidate thresholds. The guarantee follows from the fixed-sequence testing structure: an unsafe threshold can be reached only if the first unsafe candidate is incorrectly certified.

3.5. Statistical and Practical Remarks

3.5.1. Interpretation of the Guarantee

The probability in Theorem 2 is taken over the random calibration sample. The quantity $R(\hat{t})$ is the population error rate among future observations selected by the data-

dependent threshold $\hat{t}$. Thus, the result is a high-probability statement about the population selective risk, rather than a guarantee only on the empirical calibration error.

3.5.2. No Calibration Requirement on the Uncertainty Score

The validity result does not require U itself to represent a calibrated probability of error. The score only determines an ordered collection of nested selection sets. Its discriminative quality affects the achievable acceptance probability, but not the validity of the risk certificate.

3.5.3. Choice of Confidence Bound

The Hoeffding bound requires only bounded binary errors and is available in closed form. The Clopper–Pearson bound exploits the exact conditional binomial model and is often sharper when the selected sample contains few errors. Both bounds can therefore be used within the same fixed-sequence procedure.

3.5.4. Construction and Ordering of the Threshold Grid

The candidate grid must be chosen independently of the calibration sample. Scanning all unique uncertainty values obtained from that same sample is not covered by Theorem 2. When a natural score range is unavailable, one may use an independent pilot sample to construct a quantile-based grid and reserve a disjoint sample for calibration.

In practice, we recommend constructing the grid from quantiles of the uncertainty scores on an independent pilot set, which adapts naturally to skewed or model-dependent score scales. Candidates should then be ordered from lower to higher uncertainty, corresponding to increasingly permissive acceptance sets. A moderately sized grid is generally preferable to an unnecessarily dense one: adding many closely spaced candidates provides little practical resolution while creating additional opportunities for fixed-sequence testing to stop because of finite-sample fluctuations. If the pilot data suggest pronounced non-monotonicity of selective risk, one may use the pilot set to construct a coarser ascending grid that avoids unstable regions, provided that all such design choices are frozen before the calibration sample is examined. When recovering potentially safe thresholds beyond an intermediate unsafe region is operationally important, a simultaneous multiplicity-controlled procedure may be preferable to fixed-sequence testing.

3.5.5. Conservativeness of Fixed-Sequence Testing

The fixed-sequence rule stops at the first uncertified candidate. Consequently, if the population risk is non-monotone, a later threshold could, in principle, be safe even though it is never examined after an earlier failure. This is the price paid for obtaining the simple level- δ guarantee without a Bonferroni correction over all K candidates. The ordering of candidate thresholds therefore affects statistical power, although not the validity statement itself. This trade-off motivates using an independent pilot sample to design a practically meaningful candidate sequence and, in settings with substantial non-monotonicity, to consider simultaneous certification when access to later safe regions is important.

3.5.6. Unattainable Risk Targets

If CIC returns NULL, the calibration data provide insufficient evidence to certify even the most conservative candidate at risk level α and confidence level $1 - \delta$. This does not imply that no population threshold satisfying $R(t) \leq \alpha$ exists; it only states that such a threshold cannot be certified by the specified candidate set and available calibration sample.

4. Experiments

We conduct two complementary sets of experiments. First, a controlled simulation study evaluates the finite-sample statistical properties of CIC when the population selective risk is known exactly. Second, we instantiate CIC for selective question answering with large language models (LLMs) and examine its empirical behavior across models, datasets, target risk levels, and uncertainty estimators. We additionally examine the sensitivity of fixed-sequence calibration to the resolution of the candidate threshold grid in a non-monotone setting.

4.1. Simulation Study

4.1.1. Data-Generating Processes

The simulation is designed to answer three statistical questions: (i) whether the probability of returning an unsafe threshold is controlled at the nominal confidence level; (ii) how much conservativeness is introduced by different confidence bounds and multiple-testing strategies; and (iii) how fixed-sequence calibration behaves when the population selective risk is not monotone in the threshold. We further examine (iv) how the resolution of the candidate threshold grid affects the efficiency of fixed-sequence calibration under non-monotone risk.

For every simulated observation, we first sample an uncertainty score

$$U \sim \text{Uniform}(0, 1),$$

and then generate the binary error indicator according to

$$E \mid U = u \sim \text{Bernoulli}(p(u)).$$

We consider the following three scenarios:

$$p_M(u) = 0.02 + 0.30u, \quad (19)$$

$$p_{NM}(u) = \text{clip}(0.03 + 0.18u + 0.10 \sin(8\pi u), 0.001, 0.5), \quad (20)$$

$$p_{NF}(u) = 0.14. \quad (21)$$

Scenario M produces an approximately increasing selective-risk curve. Scenario NM introduces local non-monotonicity, allowing us to evaluate the fixed-sequence procedure without relying on a monotonic-risk assumption. Scenario NF has no feasible candidate at the target risk level considered below.

We set the target risk to $\alpha = 0.10$ and the significance level to $\delta = 0.05$. Unless otherwise specified, each calibration sample contains $N = 5000$ observations. We use the pre-specified candidate grid

$$\mathcal{T} = \{0.20, 0.22, \dots, 1.00\},$$

which contains $K = 41$ thresholds and is independent of all simulated calibration samples. Each configuration is repeated for 2000 independent Monte Carlo replications.

Because the data-generating distribution is known, the population selective risk $R(t)$ can be evaluated numerically to arbitrary precision. We therefore define the violation probability as the fraction of Monte Carlo replications for which the procedure returns a non-null threshold satisfying

$$R(\hat{t}) > \alpha.$$

The mean acceptance probability is $\mathbb{E}[\pi(\hat{t})]$, with a null decision assigned zero acceptance. We additionally report the probability that a procedure returns NULL.

4.1.2. Compared Calibration Procedures

We compare five procedures. Naive selects the largest threshold whose empirical selective risk does not exceed α , without uncertainty correction. Bonf-H and Bonf-CP construct simultaneous Hoeffding and Clopper–Pearson bounds, respectively, using significance level δ/K for each candidate, and then select the largest certified threshold. CIC-H and CIC-CP use the proposed fixed-sequence procedure with significance level δ at each step and stop at the first uncertified candidate.

4.1.3. Finite-Sample Risk Control

Table 1 highlights several features of the calibration problem. The naive empirical procedure substantially exceeds the nominal violation probability in both feasible scenarios, demonstrating that selecting a threshold from the same finite sample without uncertainty correction can lead to severe selection bias. In contrast, all confidence-bound-based procedures remain below the target failure probability $\delta = 0.05$.

Table 1. Finite-sample simulation at $\alpha = 0.10$, $\delta = 0.05$, and $N = 5000$. Violation is the Monte Carlo probability that a returned threshold has true population risk above α . Acceptance is the mean population acceptance probability, with NULL assigned zero acceptance. Null is the probability of returning no certified threshold. In the non-monotone scenario, Bonf-CP can attain higher acceptance than CIC-CP because simultaneous certification can reach a later safe region, whereas fixed-sequence CIC stops at the first uncertified candidate.

Scenario	Method	Violation	Acceptance	Null
M: Monotone	Naive	0.421	0.522	0.000
	Bonf-H	0.000	0.039	0.842
	Bonf-CP	0.002	0.387	0.000
	CIC-H	0.000	0.302	0.062
	CIC-CP	0.028	0.452	0.000
NM: Non-monotone	Naive	0.442	0.692	0.000
	Bonf-H	0.000	0.015	0.948
	Bonf-CP	0.002	0.473	0.011
	CIC-H	0.000	0.054	0.824
	CIC-CP	0.021	0.424	0.030

The Clopper–Pearson variants are substantially less conservative than their Hoeffding counterparts. In Scenario M, CIC-CP attains a mean acceptance probability of 0.452, compared with 0.387 for Bonf-CP, while its violation probability remains below 0.05. This illustrates the gain obtained by fixed-sequence testing relative to simultaneous Bonferroni correction when the threshold ordering is favorable.

Scenario NM also illustrates an important limitation of fixed-sequence calibration. Because testing terminates at the first uncertified candidate, CIC may fail to reach a later safe region when the population risk is locally non-monotone. Consequently, Bonf-CP attains greater acceptance than CIC-CP in this scenario despite using the more conservative multiplicity correction. This behavior is consistent with the theoretical discussion in Section 3.4: fixed-sequence testing guarantees validity without requiring risk monotonicity, but non-monotonicity may reduce statistical power.

Scenario NF is retained as a control case but omitted from Table 1 for compactness. In this setting, no candidate satisfies the target population risk. All five procedures return NULL in every Monte Carlo replication, confirming that CIC does not force deployment when the requested risk level cannot be certified within the candidate family.

4.1.4. Effect of Calibration Sample Size

We next vary the calibration size in Scenario M while keeping $\alpha = 0.10$, $\delta = 0.05$, and the candidate grid fixed. Since the Clopper–Pearson procedures are substantially more informative in the low-error regime, Table 2 focuses on Bonf-CP and CIC-CP.

Table 2. Effect of calibration size in the monotone simulation scenario. Each configuration uses 2000 Monte Carlo replications.

<i>N</i>	Method	Violation	Acceptance	Null
1000	Bonf-CP	0.000	0.109	0.633
	CIC-CP	0.026	0.293	0.175
2500	Bonf-CP	0.002	0.296	0.102
	CIC-CP	0.035	0.414	0.005
5000	Bonf-CP	0.001	0.389	0.000
	CIC-CP	0.034	0.454	0.000

As expected, larger calibration samples narrow the confidence bounds and increase the probability of certifying more permissive thresholds. The violation probability of CIC-CP remains below the nominal 0.05 level at every sample size, while its acceptance probability increases from 0.293 at $N = 1000$ to 0.454 at $N = 5000$. The advantage over Bonf-CP is largest in the smaller-sample regime, where multiplicity correction has the greatest effect on statistical power. In this favorable simulation regime, a calibration sample on the order of 10^3 observations already yields a non-null CIC-CP policy in most replications, although substantially larger samples improve acceptance. This should be interpreted as an empirical rule of thumb rather than a universal minimum: the required sample size depends on the target risk level, underlying model error rate, uncertainty ranking quality, and desired acceptance rate.

The violation probabilities of the confidence-based procedures are often below the nominal level of $\delta = 0.05$. This conservativeness is expected from the finite-sample confidence construction, the discreteness of the binomial error count, and, for CIC, the stopping structure of fixed-sequence testing. The gap between the observed and nominal violation levels suggests potential room for more statistically efficient procedures; for example, through sharper finite-sample bounds or alternative sequential testing constructions. Such extensions would need to preserve the same finite-sample validity rather than simply relax the confidence level.

4.1.5. Sensitivity to Candidate-Grid Resolution

To assess whether the efficiency of CIC is sensitive to an unnecessarily dense candidate sequence, we repeat Scenario NM with CIC-CP using four pre-specified ascending grids containing $K \in \{11, 21, 41, 81\}$ thresholds between 0.20 and 1.00. Because $U \sim \text{Uniform}(0, 1)$ in this simulation, these equally spaced thresholds also correspond to equally spaced population quantiles. All other settings remain unchanged: $N = 5000$, $\alpha = 0.10$, and $\delta = 0.05$, with 2000 Monte Carlo replications. Each grid is fixed before the calibration sample is generated. The results are reported in Table 3.

Table 3. Sensitivity of CIC-CP to candidate-grid resolution in Scenario NM.

<i>K</i>	Violation	Acceptance	Null
11	0.028	0.396	0.027
21	0.012	0.420	0.027
41	0.020	0.422	0.027
81	0.036	0.423	0.027

The results indicate that grid resolution mainly affects efficiency rather than validity. A very coarse grid ($K = 11$) loses acceptance because it provides only a small number of attainable operating points. Increasing the resolution to $K = 21$ substantially recovers this loss, whereas further increasing the grid to $K = 41$ or 81 yields only marginal additional acceptance. All observed violation probabilities remain below the nominal level. This supports the practical use of a moderately sized, ascending candidate grid: sufficient resolution is needed to avoid discretization loss, but an extremely dense grid need not provide a meaningful efficiency gain.

4.2. Selective Answering with Large Language Models

4.2.1. Models and Benchmarks

We evaluate CIC on seven instruction-tuned LLMs: Qwen2.5-3B-Instruct, Qwen2.5-7B-Instruct, Qwen2.5-14B-Instruct, OpenChat-3.5 [22], Vicuna-7B-v1.5, Vicuna-13B-v1.5 [23], and LLaMA-3.1-8B-Instruct. The models span different families and parameter scales, providing a range of base error rates and uncertainty-quality regimes.

We consider CommonsenseQA [24] and TriviaQA [25]. CommonsenseQA is treated as a closed-ended prediction problem, where correctness is determined by agreement with the ground-truth answer option. TriviaQA is an open-ended factual QA task. Following prior work [5,7], we compute cosine similarity between fixed sentence embeddings of the generated and reference answers and classify a response as correct when the similarity is at least 0.6.

4.2.2. Generation and Uncertainty Estimation

Semantic entropy [6] is used as the primary uncertainty score. For each input, we generate one response for correctness evaluation and draw 10 auxiliary stochastic responses for uncertainty estimation. The auxiliary samples use temperature 1.0, top- $p = 0.9$, and a maximum generation length of 64 tokens. Semantic entropy is computed from the resulting semantic clusters following the original procedure. The same generation and uncertainty settings are used across all models.

Model outputs, binary error labels, and uncertainty scores are generated once for each question and then held fixed across repeated data partitions. Hence, the repeated splits measure calibration variability rather than additional generation variability.

4.2.3. Pilot, Calibration, and Test Splits

For each model–dataset pair, we randomly partition the available questions at the question level into three disjoint subsets: 10% pilot, 45% calibration, and 45% test. The pilot subset is used only to construct the candidate threshold grid. Specifically, we compute empirical uncertainty quantiles at

$$q \in \{0.20, 0.25, \dots, 1.00\},$$

remove duplicate thresholds if necessary, and order the resulting candidates from low to high uncertainty. No calibration or test observation is used to construct the candidate set.

The calibration subset is then used by CIC to select the deployment threshold, and the test subset is used only for evaluation. This separation ensures that the candidate threshold sequence is independent of the calibration sample, as required by Theorem 2. We set $\delta = 0.05$ throughout and consider

$$\alpha \in \{0.05, 0.10, 0.15, 0.20, 0.25\}.$$

Each experiment is repeated over 100 random partitions.

The use of an independent pilot subset serves two purposes. First, it prevents the candidate thresholds from being adapted to the same calibration observations used for statistical certification. Second, quantile-based construction provides a simple way to obtain a moderately sized grid whose candidate acceptance sets are reasonably distributed across the observed uncertainty-score range. All candidate thresholds and their ordering are fixed before the calibration subset is examined.

4.2.4. Evaluation Metrics

We avoid the term “power” for deployment coverage because it has a distinct meaning in statistical hypothesis testing. Instead, we report the following:

- Selective risk (SR): The proportion of incorrect responses among accepted test outputs;
- Acceptance rate (AR): The proportion of test questions for which the calibrated policy returns an answer;
- Null rate: The proportion of random partitions in which no threshold can be certified.

For the SR and AR, means are computed over non-null trials unless stated otherwise. The null rate is computed over all 100 random partitions. Importantly, the test SR is an empirical estimate of $R(\hat{t})$ and should not be interpreted as the exact violation event appearing in Theorem 2; the simulation study provides the direct finite-sample validation because the population risk is known there.

4.3. LLM Results

4.3.1. Model-Wise Selective Risk and Acceptance

Table 4 reports representative model-wise results at $\alpha = 0.10$. Across both datasets, the Clopper–Pearson implementation generally retains more responses than the Hoeffding implementation while keeping the mean selective risk near or below the prescribed target. The difference is particularly visible for models with higher base error rates, where Hoeffding bounds may be too wide to certify a useful threshold.

Table 4. Model-wise selective risk (SR), acceptance rate (AR), and null rate at $\alpha = 0.10$. Results are averaged over 100 random partitions.

Model	Bound	CommonsenseQA			TriviaQA		
		SR	AR	Null	SR	AR	Null
Qwen2.5-3B-Instruct	H	0.074	0.502	0.00	0.064	0.381	0.00
	CP	0.090	0.623	0.00	0.086	0.517	0.00
Qwen2.5-7B-Instruct	H	0.079	0.720	0.00	0.077	0.741	0.00
	CP	0.092	0.766	0.00	0.091	0.833	0.00
OpenChat-3.5	H	0.079	0.802	0.00	0.078	0.889	0.00
	CP	0.092	0.839	0.00	0.092	0.929	0.00
Vicuna-7B-v1.5	H	–	–	1.00	0.065	0.425	0.00
	CP	0.080	0.233	0.66	0.086	0.567	0.00
LLaMA-3.1-8B-Instruct	H	0.074	0.551	0.00	0.078	0.904	0.00
	CP	0.090	0.608	0.00	0.090	0.937	0.00
Vicuna-13B-v1.5	H	–	–	1.00	0.077	0.835	0.00
	CP	0.071	0.232	0.00	0.090	0.891	0.00
Qwen2.5-14B-Instruct	H	0.080	0.833	0.00	0.079	0.970	0.00
	CP	0.092	0.865	0.00	0.088	0.999	0.00

The results also illustrate the role of the underlying model–uncertainty pair. For relatively strong models such as Qwen2.5-14B-Instruct, a large fraction of responses can be retained while satisfying a strict risk target. For weaker configurations, such as Vicuna on

CommonsenseQA, the calibration procedure may return NULL because the available data cannot certify the requested risk. Such cases should be viewed as an intended statistical outcome rather than an algorithmic failure.

4.3.2. Behavior Across Target Risk Levels

Table 5 summarizes the behavior across the complete set of target risk levels, pooled over the seven models and two datasets. As α increases, confidence-based calibration permits more permissive thresholds and therefore increases the acceptance rate. The mean selective risk correspondingly moves toward the prescribed target or saturates at the base-model error rate.

Table 5. Pooled selective risk, acceptance rate, and null rate across the seven models and two QA benchmarks. The SR and AR are averaged over non-null trials.

α	Bound	SR	AR	Null
0.05	H	0.028	0.310	0.67
	CP	0.044	0.390	0.24
0.10	H	0.078	0.660	0.13
	CP	0.090	0.750	0.02
0.15	H	0.124	0.830	0.05
	CP	0.138	0.870	0.00
0.20	H	0.164	0.910	0.01
	CP	0.178	0.940	0.00
0.25	H	0.190	0.960	0.00
	CP	0.201	0.970	0.00

The pooled results display the expected risk–acceptance trade-off. At $\alpha = 0.05$, both confidence bounds are necessarily conservative and null decisions occur frequently, whereas for moderate risk levels, most model–dataset combinations admit a non-null certified policy. The CP variant consistently achieves greater acceptance, reflecting the sharper finite-sample binomial bound observed in the simulation study.

4.3.3. Comparison with Uncorrected and Simultaneous Calibration

To connect the LLM application with the statistical simulation, we additionally compare CIC against the naive empirical rule and Bonferroni-corrected simultaneous certification at $\alpha = 0.10$. Table 6 reports results pooled across all model–dataset combinations.

Table 6. Comparison of calibration strategies at $\alpha = 0.10$, pooled across the LLM experiments. The SR and AR are averaged over non-null trials.

Method	SR	AR	Null
Naive	0.112	0.820	0.00
Bonf-H	0.057	0.390	0.41
Bonf-CP	0.073	0.550	0.12
CIC-H	0.078	0.660	0.13
CIC-CP	0.090	0.750	0.02

The naive rule achieves the highest acceptance but exceeds the target risk on average, illustrating the danger of ignoring threshold-selection uncertainty. Bonferroni correction is substantially more conservative. CIC, particularly CIC-CP, occupies an intermediate regime in which the observed selective risk remains compatible with the target while substantially more predictions are retained.

4.3.4. Robustness to the Uncertainty Score

Although semantic entropy is used in the main experiments, CIC is statistically agnostic to how the uncertainty ranking is obtained. We therefore repeat the $\alpha = 0.10$ experiment using semantic entropy (SE), shift attention to relevance (SAR) [5], and word-sequence entropy (WSE) [7]. To keep this ablation computationally manageable, we use Qwen2.5-3B-Instruct, Qwen2.5-7B-Instruct, and LLaMA-3.1-8B-Instruct and report averages over the three models.

For uncertainty measures whose numerical range is not fixed in advance, the candidate thresholds are constructed from the independent pilot subset using the same quantile grid as in the main experiment. CIC-CP is used for certification. The results are reported in Table 7.

Table 7. Effect of the uncertainty estimator at $\alpha = 0.10$ using CIC-CP. Results are averaged over three representative LLMs.

UQ Score	CommonsenseQA		TriviaQA	
	SR	AR	SR	AR
SE	0.091	0.640	0.090	0.810
SAR	0.092	0.670	0.091	0.830
WSE	0.089	0.690	0.088	0.850

Across the three uncertainty measures, the selective risk remains close to the same prescribed level, while the acceptance rate varies more noticeably. This is the expected behavior of the proposed framework: statistical validity is supplied by the calibration layer, whereas the quality of the uncertainty ranking mainly determines efficiency.

4.4. Summary of Empirical Findings

The numerical studies support four conclusions. First, threshold selection based only on empirical selective risk can substantially exceed the intended population risk, whereas confidence-bound-based calibration controls this selection effect. Second, fixed-sequence CIC can provide substantially greater acceptance than a Bonferroni simultaneous correction when the candidate ordering is favorable. Third, the exact Clopper–Pearson implementation is generally more statistically efficient than the Hoeffding implementation in the low-error regimes encountered in selective QA. Fourth, the same calibration mechanism can be applied to different uncertainty scores: changing the uncertainty estimator primarily changes the attainable acceptance rate rather than the form of the statistical guarantee.

The candidate-grid sensitivity study further suggests that an extremely dense grid is not necessary in the examined non-monotone setting: moving from a coarse grid to a moderate resolution substantially improves acceptance, while further densification provides only a small additional gain.

At the same time, the non-monotone simulation clarifies an important limitation. Fixed-sequence testing may stop before reaching a later safe threshold when the population selective risk fluctuates across the candidate sequence. Thus, CIC is best interpreted as a finite-sample-valid calibration procedure whose efficiency depends jointly on the quality of the uncertainty ordering, the calibration sample size, and the structure of the candidate threshold sequence. The observed conservativeness also leaves room for future work on sharper confidence constructions and more efficient sequential certification procedures, provided that finite-sample risk validity is retained.

5. Conclusions

This paper introduced confidence-interval calibration (CIC), a finite-sample procedure for converting uncertainty scores into risk-controlled selective decision rules. Rather than requiring the uncertainty score itself to represent a calibrated probability of error, CIC uses the score only to define an ordered family of candidate selection policies. One-sided Hoeffding or exact Clopper–Pearson upper confidence bounds are then combined with fixed-sequence testing to certify increasingly permissive thresholds. Under independent and identically distributed calibration and deployment samples, the resulting procedure guarantees, with a probability of at least $1 - \delta$, that any returned non-null threshold satisfies the prescribed population acceptance-conditioned risk level. The guarantee applies to the data-dependent threshold returned by the procedure and does not require monotonicity of the population selective risk across candidate thresholds.

The numerical studies examine both the statistical properties of the calibration procedure and its use in large language model (LLM) selective answering. Controlled simulations show why empirical threshold selection without uncertainty correction can lead to substantial risk violations, and illustrate the trade-off between simultaneous Bonferroni correction and fixed-sequence certification. In particular, the Clopper–Pearson implementation can retain substantially greater acceptance than more conservative alternatives when the threshold ordering is informative. The simulations also clarify the behavior under non-monotone selective risk: the validity guarantee remains intact, although fixed-sequence stopping may reduce acceptance by preventing the procedure from reaching later safe thresholds. The additional candidate-grid sensitivity study further indicates that, in the examined non-monotone setting, a moderately resolved grid can recover most of the attainable acceptance while further densification provides only limited additional benefit. Experiments on CommonsenseQA and TriviaQA further demonstrate that CIC can be applied across different LLMs and uncertainty estimators, with statistical calibration governing reliability while the quality of the underlying uncertainty ranking primarily determines the attainable acceptance rate. Together, these results suggest that CIC is most naturally viewed as a statistical certification layer placed on top of an existing model–uncertainty pair, rather than as a replacement for improving the predictive model or uncertainty estimator itself.

Several limitations clarify both the scope of the present guarantee and the settings in which additional methodology is needed. The calibration sample and deployment distribution are assumed to be identically distributed, and the candidate threshold sequence must be fixed independently of the calibration sample. Moreover, fixed-sequence testing trades statistical efficiency for a simple finite-sample guarantee, and may be conservative when the population risk varies irregularly across thresholds. In particular, CIC does not attempt to identify the globally most permissive safe threshold when the risk curve is non-monotone; it guarantees validity for the threshold reached through the pre-specified testing sequence. In open-ended applications, the validity of the resulting risk statement also depends on the correctness criterion used to define the binary error labels.

The empirical violation probabilities observed in our simulations are often below the nominal level δ , reflecting conservativeness from finite-sample confidence bounds, discrete error counts, and the fixed-sequence stopping rule. This gap also points to a natural direction for improving statistical efficiency without weakening the target guarantee. Future work may therefore consider calibration under distribution shift, data-adaptive threshold construction with valid post-selection inference, sharper sequential or simultaneous confidence procedures, and extensions to more general losses, multiple risk constraints, and structured selective decision problems. An additional direction would be to develop less conservative procedures that retain finite-sample validity while making better use of later safe regions under non-monotone selective risk.

Author Contributions: Conceptualization, S.D. and H.S.; methodology, S.D.; software, S.D.; validation, S.D. and X.H.; formal analysis, S.D.; investigation, S.D. and X.H.; data curation, S.D.; visualization, S.D.; writing—original draft preparation, S.D.; writing—review and editing, S.D., X.H., and H.S.; supervision, H.S.; project administration, H.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets used in this study are publicly available. CommonsenseQA is available at https://huggingface.co/datasets/tau/commonsense_qa (accessed on 15 March 2026), and TriviaQA is available at https://huggingface.co/datasets/mandarjoshi/trivia_qa (accessed on 15 March 2026). No new datasets were created in this study.

Acknowledgments: During the preparation of this manuscript, the authors used OpenAI ChatGPT (GPT-5.6 Sol) for language editing and manuscript organization. The authors reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Wang, S.; Xu, T.; Li, H.; Zhang, C.; Liang, J.; Tang, J.; Yu, P.S.; Wen, Q. Large language models for education: A survey and outlook. *IEEE Signal Process. Mag.* **2025**, *42*, 51–63.
2. Jiang, J.; Wang, F.; Shen, J.; Kim, S.; Kim, S. A survey on large language models for code generation. *ACM Trans. Softw. Eng. Methodol.* **2026**, *35*, 58.
3. Gawlikowski, J.; Tassi, C.R.N.; Ali, M.; Lee, J.; Humt, M.; Feng, J.; Kruspe, A.; Triebel, R.; Jung, P.; Roscher, R.; et al. A survey of uncertainty in deep neural networks. *Artif. Intell. Rev.* **2023**, *56*, 1513–1589.
4. Geifman, Y.; El-Yaniv, R. Selective classification for deep neural networks. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 4878–4887.
5. Duan, J.; Cheng, H.; Wang, S.; Zavalny, A.; Wang, C.; Xu, R.; Kailkhura, B.; Xu, K. Shifting attention to relevance: Towards the predictive uncertainty quantification of free-form large language models. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Bangkok, Thailand, 11–16 August 2024; pp. 5050–5063.
6. Kuhn, L.; Gal, Y.; Farquhar, S. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In Proceedings of the Eleventh International Conference on Learning Representations, Kigali, Rwanda, 1–5 May 2023.
7. Wang, Z.; Duan, J.; Yuan, C.; Chen, Q.; Chen, T.; Zhang, Y.; Wang, R.; Shi, X.; Xu, K. Word-sequence entropy: Towards uncertainty estimation in free-form medical question answering applications and beyond. *Eng. Appl. Artif. Intell.* **2025**, *139*, 109553.
8. Wang, Z.; Duan, J.; Cheng, L.; Zhang, Y.; Wang, Q.; Shi, X.; Xu, K.; Shen, H.T.; Zhu, X. ConU: Conformal uncertainty in large language models with correctness coverage guarantees. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, FL, USA, 12–16 November 2024; pp. 6886–6898.
9. Guo, C.; Pleiss, G.; Sun, Y.; Weinberger, K.Q. On calibration of modern neural networks. In *Proceedings of the International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017*; PMLR: Cambridge, MA, USA, 2017; pp. 1321–1330.
10. Gui, Y.; Jin, Y.; Ren, Z. Conformal alignment: Knowing when to trust foundation models with guarantees. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 73884–73919.
11. Wang, Z.; Duan, J.; Wang, Q.; Zhu, X.; Chen, T.; Shi, X.; Xu, K. COIN: Uncertainty-Guarding Selective Question Answering for Foundation Models with Provable Risk Guarantees. *Proc. AAAI Conf. Artif. Intell.* **2026**, *40*, 33764–33772.
12. Wang, Z.; Aniri, C.; Chen, T.; Zhang, Y.; Shen, H.T.; Shi, X.; Xu, K. LEC: Linear Expectation Constraints for Selection-Conditioned Risk Control in Selective Prediction and Routing Systems. In Proceedings of the Forty-Third International Conference on Machine Learning, Seoul, Republic of Korea, 6–11 July 2026; pp. 124814–124840.
13. Angelopoulos, A.N.; Bates, S.; Jordan, M.; Malik, J. Uncertainty sets for image classifiers using conformal prediction. In Proceedings of the International Conference on Learning Representations, Virtual, 3–7 May 2021.
14. Angelopoulos, A.N.; Bates, S. Conformal prediction: A gentle introduction. *Found. Trends Mach. Learn.* **2023**, *16*, 494–591.
15. Wang, Z.; Wang, Q.; Zhang, Y.; Chen, T.; Zhu, X.; Shi, X.; Xu, K. SConU: Selective conformal uncertainty in large language models. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Vienna, Austria, 27 July–1 August 2025; pp. 19052–19075.

16. Wang, Q.; Geng, T.; Wang, Z.; Wang, T.; Fu, B.; Zheng, F. Sample then Identify: A General Framework for Risk Control and Assessment in Multimodal Large Language Models. In Proceedings of the Thirteenth International Conference on Learning Representations, Singapore, 24–28 April 2025.
17. Clopper, C.J.; Pearson, E.S. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* **1934**, *26*, 404–413.
18. Campos, M.; Farinhas, A.; Zerva, C.; Figueiredo, M.A.; Martins, A.F. Conformal prediction for natural language processing: A survey. *Trans. Assoc. Comput. Linguist.* **2024**, *12*, 1497–1516.
19. Quach, V.; Fisch, A.; Schuster, T.; Yala, A.; Sohn, J.H.; Jaakkola, T.S.; Barzilay, R. Conformal language modeling. In Proceedings of the Twelfth International Conference on Learning Representations, Vienna, Austria, 7–11 May 2024.
20. Mohri, C.; Hashimoto, T. Language models with conformal factuality guarantees. In Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria, 21–27 July 2024; pp. 36029–36047.
21. Geifman, Y.; El-Yaniv, R. SelectiveNet: A deep neural network with an integrated reject option. In *Proceedings of the International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019*; PMLR: Cambridge, MA, USA, 2019; pp. 2151–2159.
22. Wang, G.; Cheng, S.; Zhan, X.; Li, X.; Song, S.; Liu, Y. OpenChat: Advancing open-source language models with mixed-quality data. In Proceedings of the International Conference on Learning Representations, Vienna, Austria, 7–11 May 2024; Volume 2024, pp. 57021–57040.
23. Zheng, L.; Chiang, W.L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.; et al. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 46595–46623.
24. Talmor, A.; Herzig, J.; Lourie, N.; Berant, J. CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, MN, USA, 2–7 June 2019; pp. 4149–4158.
25. Joshi, M.; Choi, E.; Weld, D.; Zettlemoyer, L. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Vancouver, BC, Canada, 30 July–4 August 2017; pp. 1601–1611.

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 MDPI and/or the editor(s). MDPI 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.