COIN: Uncertainty-Guarding Selective Question Answering for Foundation Models with Provable Risk Guarantees

Zhiyuan Wang[1], Jinhao Duan[2], Qingni Wang[1], Xiaofeng Zhu[1], Tianlong Chen[3], Xiaoshuang Shi[1]*, Kaidi Xu[2]*

[1]University of Electronic Science and Technology of China
[2]Drexel University
[3]University of North Carolina at Chapel Hill

Abstract

Uncertainty quantification (UQ) in foundation models is crucial for identifying and mitigating hallucinations in automatically generated text. However, heuristic UQ approaches lack statistical guarantees for key metrics such as the false discovery rate (FDR) in selective prediction tasks. Previous research adopts the split conformal prediction (SCP) framework to ensure desired coverage of admissible answers by constructing data-driven prediction sets, yet these sets typically contain incorrect candidates, undermining their practical effectiveness. To address this, we introduce COIN, an uncertainty-guarding selection framework that calibrates statistically valid uncertainty thresholds to filter a single generated answer per question under user-specified FDR constraints. COIN estimates the empirical error rate on the calibration set and applies confidence interval methods such as Clopper–Pearson to establish a high-probability upper bound on the true error rate (i.e., FDR). This enables the selection of the largest threshold that ensures FDR control on test data while significantly increasing sample retention. We demonstrate COIN's robustness in risk control, strong test-time power in retaining admissible answers, and predictive efficiency under limited calibration data across both general and multimodal text generation tasks. Furthermore, we show that employing alternative UQ and upper bound construction strategies can further boost COIN's power performance, which underscores its extensibility and adaptability to diverse application scenarios.

Code —
https://github.com/Zhiyuan-GG/COIN-AAAI-2026

Introduction

Recent advancements in foundation models, especially large language models (LLMs) and large vision-language models (LVLMs) (Jaech et al. 2024; Guo et al. 2025), have remarkably propelled progress in a wide range of downstream tasks such as question answering (QA) (Singhal et al. 2025). Nevertheless, these models remain susceptible to trustworthiness issues like hallucination, prone to responding with plausible yet inadmissible information (Huang et al. 2025; Duan et al. 2025). Such deficiencies compromise their reliable deployment in risk-sensitive domains (Zheng et al. 2025; Pennydimri et al. 2025). A prevalent mitigation strategy is to estimate model uncertainty during communication, thereby enabling selective abstention when predictions exhibit high uncertainty (Wang et al. 2025b).

Although heuristic uncertainty quantification (UQ) methods perform well in practice, they offer no statistical guarantees for controlling critical metrics like the false discovery rate (FDR) in selective generation scenarios (Chen et al. 2023; Tayebati et al. 2025). Recent studies have adapted the split conformal prediction (SCP) framework (Angelopoulos and Bates 2021), which converts heuristic uncertainty estimates from arbitrary models to statistically rigorous ones by constructing prediction sets, to QA tasks, ensuring coverage of admissible responses at a user-specified risk level (Wang et al. 2024; Ye et al. 2024; Quach et al. 2024; Wang et al. 2025a). However, these prediction sets typically include inadmissible candidates, which severely limits their practical utility (Cresswell et al. 2025). Consequently, determining statistically valid uncertainty thresholds that rigorously control FDR for selective prediction tasks, requiring a single definitive answer per input, remains an open challenge.

A previous work employs the conformal alignment (CA) framework (Gui, Jin, and Ren 2024), which first trains an alignment predictor and then applies the conformal p-value framework to filter new test samples (Jin and Candès 2023; Huang, Lala, and Jha 2024). An answer is retained when its corresponding predicted alignment score surpasses a data-dependent threshold. This conformalized selection mechanism ensures FDR control among the selected QA samples. However, CA requires traversing the entire calibration set to compute the p-value for each test sample, and depends on the Benjamini–Hochberg (BH) procedure (Benjamini and Hochberg 1995) to determine the p-value threshold, making it highly time-consuming for large-scale QA tasks and causing unnecessary rejection of many admissible samples.

In this paper, we introduce an uncertainty-guarding selective prediction framework, termed **COIN**, with three stages comprising **C**alibration set error analysis, upper b**O**und construct**I**on, and threshold selectio**N**, as illustrated in Figure 1. Each stage is implemented as a modular component to facilitate separate optimization and flexible substitution. Following the principles of SCP-based frameworks, COIN begins by holding out a dedicated calibration set for threshold calibration. Given a candidate threshold, we quantify the

*Corresponding Authors.
Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

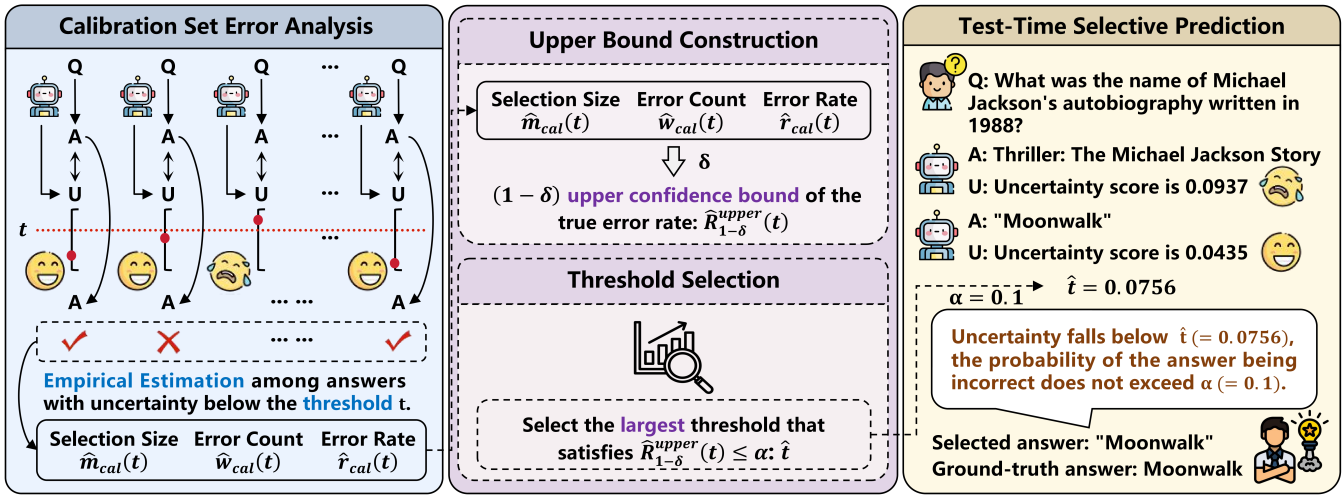


Figure 1: An overview of the three-stage COIN framework and an illustration of the test-time selective prediction.

uncertainty for each calibration data, retaining only those samples whose uncertainty scores fall below the threshold, i.e., treating them as trustworthy. Subsequently, we estimate the empirical error rate over this uncertainty-filtered calibration subset. To rigorously constrain selection risks based on test statistics from the in-distribution calibration data, COIN proceeds to construct a high-probability upper confidence bound of the true failure/error rate, computed from observations of the retained subset. Finally, we identify the largest threshold whose corresponding upper bound remains below the risk level, thereby providing robust control over the FDR on test data while maximizing sample retention.

Specifically, we record the size of the uncertainty-filtered calibration subset and the number of failure cases (inadmissible answers) in the first stage, from which the empirical failure rate is computed. These statistics serve as the foundation for the second stage, where we apply the one-sided Clopper–Pearson interval (Clopper and Pearson 1934) to derive an exact, high-probability upper bound on the true failure rate. Owing to COIN's modular and flexible design, alternative concentration inequalities such as Hoeffding's inequality (Hoeffding 1994; Bentkus 2004; Bates et al. 2021) can also be employed to construct upper confidence bounds. While these alternatives are not exact, they significantly improve computational efficiency in large-scale QA scenarios while maintaining valid statistical guarantees. With the high-probability upper bound in place, the third stage of COIN focuses on threshold selection. By progressively increasing the uncertainty threshold, we aim to determine the maximum value such that the associated upper bound on the failure rate does not exceed a user-specified risk level. This choice ensures rigorous control over test-time risk while maximizing the number of selected admissible answers, striking a trade-off between statistical rigor and predictive efficiency.

COIN is applicable to a wide range of text generation applications (**generality**). In this study, we evaluate its statistical validity on two textual QA datasets: closed-ended CommonsenseQA (Talmor et al. 2019) and open-domain TriviaQA (Joshi et al. 2017), employing five LLMs. Additionally, we also implement COIN on the open-domain multimodal MMVet dataset (Yu et al. 2024) across five LVLMs. Empirical results confirm that COIN significantly outperforms CA in retention of admissible answers (**power**) while constraining the FDR at desired risk levels (**rigor**). Moreover, owing to COIN's **customizable** structure, we leverage both white-box and black-box UQ approaches in the first stage, to obtain more accurate sample filtering, and select the optimal upper bound construction strategy in the second stage, to balance computational load and sample retention. Furthermore, we validate COIN's **robustness** and predictive **efficiency** across varying splits of calibration and test data. The main contributions of this paper are summarized as follows:

- We investigate uncertainty-guarding selective prediction for general QA tasks with user-specified FDR constraints while maximizing the retention of admissible samples, which is a previously under-explored topic.
- COIN precomputes the maximal uncertainty threshold on the calibration set for a target risk level, substantially improving test-time selection efficiency.
- Extensive evaluations on three QA datasets with different response formats and modalities demonstrate that COIN rigorously constrains the failure rate of retained samples under the condition that uncertainty falls below the calibrated threshold while covering more correct answers.

Related Work

SCP in QA tasks. Recent studies have applied SCP to natural language processing tasks (Campos et al. 2024). Two influential studies (Kumar et al. 2023; Ye et al. 2024) adapt SCP to closed-ended QA tasks, providing statistical coverage guarantees while employing the size of the prediction set as a proxy for uncertainty. In open-domain settings, several studies (Quach et al. 2024; Wang et al. 2024; Kaur et al. 2024; Wang et al. 2025a) extend SCP by introducing conformalized uncertainty criteria that construct prediction sets

with user-desired coverage of admissible responses. While promising, both lines of research suffer from selection bias within prediction sets (Jin and Candès 2023; Jin and Ren 2024). For tasks that require a single answer per question, two works (Mohri and Hashimoto 2024; Cherian, Gibbs, and Candes 2024) attempt to ensure conformal factuality by filtering out unreliable sub-claims from model outputs, but such modification may remove many valuable and accurate claims, resulting in vague or uninformative responses.

Abstention. Our problem setting falls under selective prediction, where the model is allowed to either generate an output or abstain when uncertain (Kadavath et al. 2022). Prior studies have explored UQ techniques during communication to inform users of model output reliability (Su et al. 2024; Wang et al. 2025b; Yang et al. 2025; Abbasli et al. 2025). For example, VL-Uncertainty (Zhang, Zhang, and Zheng 2024) perturbs both the input images and questions to estimate the semantic entropy (Kuhn, Gal, and Farquhar 2023). If the uncertainty score surpasses a predefined threshold, the model abstains, treating the output as a potential hallucination. However, these abstention strategies lack formal guarantees (Yadkori et al. 2024; Gui, Jin, and Ren 2024), such as controlling the FDR over selected samples.

To address this, CA (Gui, Jin, and Ren 2024) fits an alignment function and applies conformalized selection to test data. While CA constrains FDR, it requires comparing each test sample against the entire calibration set to compute p-values, and its use of the BH procedure rejects many correct answers. Instead, we follow the idea of learn then test (LTT) (Angelopoulos et al. 2021), which calibrates a candidate set of parameters on the calibration set, each offering guaranteed risk control. This enables us to select the optimal parameter (threshold) that maximizes the selection of QA samples while maintaining the desired risk level.

Methodology

Preliminaries

Problem Formulation. Let $F : \mathcal{X} \to \mathcal{Y}$ be an LLM, and let U be a scalar function that estimates the model's uncertainty for a given input. For each test question $X_{test} \in \mathcal{X}$ with its ground-truth answer $Y^*_{test} \in \mathcal{Y}$, the model generates a candidate answer $\hat{Y}_{test} \sim \mathbb{P}(Y \mid X_{test})$, and we compute the associated uncertainty score as $U(F(X_{test}))$. We consider an uncertainty-guarding deployment policy, where a candidate response is retained only if the estimated uncertainty score falls below a specified threshold t, i.e., $U(F(X_{test})) \leq t$. **Our goal** is to calibrate the threshold t in a statistically rigorous way so that the probability of accepting an incorrect answer, conditioned on a low uncertainty score (i.e., $\leq t$), remains below a desired level with high probability. Formally, for a user-specified risk level $\alpha \in (0, 1)$ and the confidence level $1 - \delta$ (e.g., 0.99), we aim to guarantee

$$\Pr\Big(\Pr\big(A(\hat{Y}_{\text{test}}, Y^*_{\text{test}}) = 0 \big| U(F(X_{\text{test}})) \leq t\big) \leq \alpha\Big) \geq 1 - \delta \tag{1}$$

where $A\left(\hat{Y}, Y^*\right) \in \{0, 1\}$ is an admission indicator function that returns 1 if the candidate response $\hat{Y}$ matches the corresponding ground truth answer Y^* and 0 otherwise. The inner probability measures the true conditional failure rate (TCFR) under the event $U(F(X)) \leq t$.

To formalize the guarantee over random data, let (X, Y^*) be drawn from a data-generating distribution $\mathcal{D}$, where $X \in \mathcal{X}$ is a random question variable and $Y^* \in \mathcal{Y}$ is the corresponding random ground-truth variable, and $\hat{Y}$ denote the model's output to X. We define a binary correctness variable $Z := \mathbf{1}\left\{A\left(\hat{Y}, Y^*\right) = 1\right\}$ indicating whether the candidate answer $\hat{Y}$ of X is admissible ($Z = 1$) or not ($Z = 0$). With this notation, the TCFR under the threshold t is defined as

$$R(t) := \mathbb{E}\left[1 - Z \mid U(F(X)) \leq t\right], \tag{2}$$

and the statistical guarantee becomes

$$\Pr\left(R(t) \leq \alpha\right) \geq 1 - \delta, \tag{3}$$

following a PAC-style risk control (Snell et al. 2023; Zollo et al. 2024; Ni et al. 2025), which forms the statistical foundation of our threshold calibration framework.

Building on prior work on selective prediction (Gui, Jin, and Ren 2024; Yadkori et al. 2024; Wang et al. 2025c) and SCP (Wang et al. 2024; Quach et al. 2024; Shahrokhi et al. 2025), where QA samples (X_i, Y_i^*) are assumed to be *independent and identically distributed* (i.i.d.) data points drawn from the underlying data-generating distribution $\mathcal{D}$ (Angelopoulos et al. 2021; Angelopoulos and Bates 2021; Angelopoulos, Barber, and Bates 2024), we next formalize the statistical structure induced by the i.i.d. data-generating process together with the uncertainty-guarding selection mechanism, which underlies the further estimation of the TCFR.

Bernoulli Structure under Conditional Selection. Given a threshold t, we define a (random) subset of selected samples as $\mathcal{S}_t := \{(X_i, Y_i^*) : U(F(X_i)) \leq t\}$, which corresponds to a subpopulation drawn from the *conditional distribution* $\mathcal{D}_t := \mathcal{D} \mid U(F(X)) \leq t$. In practice, we can compute the (random) empirical failure rate on this subset as

$$\hat{R}(t) = \frac{1}{|\mathcal{S}_t|} \sum_{(X_i, Y_i^*) \in \mathcal{S}_t} (1 - Z_i). \tag{4}$$

Given that the TCFR $R(t)$ is unknown and cannot be directly evaluated on test samples, we employ the empirical counterpart $\hat{R}(t)$ observed from the selected subset $\mathcal{S}_t$ to construct a valid upper bound of $R(t)$ that holds with high probability. To enable statistically sound estimation, we follow the standard assumption that *the original QA samples (X_i, Y_i^*) are i.i.d. data points from the data-generating distribution $\mathcal{D}$*, which is widely adopted in recent risk control frameworks of QA tasks (Kumar et al. 2023; Wang et al. 2024; Ye et al. 2024; Gui, Jin, and Ren 2024; Wang et al. 2025c; Ni et al. 2025). We then introduce the following proposition.

Proposition 1. *Let (X_i, Y_i^*) be i.i.d. samples drawn from a joint distribution $\mathcal{D}$, where $Y_i^* \sim \mathbb{P}(Y \mid X_i)$. The selection indicator is defined as $I_i := \mathbf{1}\{U(F(X_i)) \leq t\}$, where $U(F(X_i))$ is a deterministic function of X_i, and the correctness indicator is $Z_i := \mathbf{1}\left\{A\left(\hat{Y}_i, Y_i^*\right) = 1\right\}$, where $\hat{Y}_i \sim \mathbb{P}(Y \mid X_i)$, then the subset $\mathcal{S}_t$ corresponds to i.i.d.*

samples from the conditional distribution $\mathcal{D}_t$, and the correctness indicators $\{Z_i\}_{(X_i,Y_i^*)\in\mathcal{S}_t}$ are i.i.d. Bernoulli random variables with success probability

$$p_t := 1 - R(t), \tag{5}$$

where $R(t)$ is the TCFR. Equivalently, the failure indicators $W_i := 1 - Z_i \sim \text{Bernoulli}(R(t))$ are also i.i.d. Bernoulli random variables under this conditional distribution $\mathcal{D}_t$.

We provide a formal justification of Proposition 1 in the appendix. The Bernoulli property is concluded from the condition of i.i.d. data-generating and uncertainty-guarding selection, without introducing additional assumptions (Quach et al. 2024; Ni et al. 2025; Shahrokhi et al. 2025). We next leverage the empirical failure rate $\hat{R}(t)$ as a consistent and unbiased estimator of the TCFR $R(t)$ to construct a statistically valid upper confidence bound with high probability.

Risk-Constrained Threshold Calibration

Building on the established statistical structure, we present the full calibration procedure of COIN for determining a statistically rigorous uncertainty threshold, denoted as $\hat{t}$, that ensures a high-confidence guarantee on the TCFR $R(t)$. Our approach follows the standard SCP-based frameworks by reserving a held-out set of M i.i.d. calibration data points, $\mathcal{D}_{cal} = \{(x_i, y_i^*)\}_{i=1}^{M} \subset \mathcal{D}$, where $x_i \in \mathcal{X}$ and $y_i^* \in \mathcal{Y}$ represent the i-th question and the corresponding ground-truth answer, respectively (Papadopoulos et al. 2002; Angelopoulos and Bates 2021; Angelopoulos, Barber, and Bates 2024). Generally, the COIN framework consists of three stages: ❶ empirical failure estimation on the calibration set, ❷ upper confidence bound construction, and ❸ threshold selection.

Empirical Failure Estimation on the Calibration Set. In the first stage, for each question x_i, we produce a candidate answer $\hat{y}_i$ and the associated uncertainty score $U(F(x_i))$. For a given threshold t, we then define the empirical conditional failure rate (ECFR) on $\mathcal{D}_{cal}$ as:

$$\hat{r}_{cal}(t) = \hat{w}_{cal}(t) / \hat{m}_{cal}(t), \tag{6}$$

where $\hat{m}_{cal}(t)$ is the number of selected calibration samples whose uncertainty scores fall below the threshold t

$$\hat{m}_{cal}(t) = \textstyle\sum_{(x_i,y_i^*)\in\mathcal{D}_{cal}} \mathbf{1}\{U(F(x_i)) \le t\}, \tag{7}$$

and $\hat{w}_{cal}(t)$ is the number of failures observed from the selected subset of $\mathcal{D}_{cal}$

$$\hat{w}_{\text{cal}}(t) = \textstyle\sum_{(x_i,y_i^*)\in\mathcal{D}_{\text{cal}}} \mathbf{1}\big\{U(F(x_i)) \le t \wedge A(\hat{y}_i, y_i^*) = 0\big\}. \tag{8}$$

The (observation) ECFR $\hat{r}_{cal}(t)$ then serves as the foundation for constructing the upper bound of the TCFR $R(t)$.

Upper Confidence Bound Construction. To rigorously constrain the TCFR $R(t)$ (i.e., FDR), we aim to construct an upper confidence bound, denoted as $\hat{R}^{\text{upper}}_{1-\delta}$, based on the ECFR $\hat{r}_{cal}(t)$ estimated from the calibration set, such that

$$\Pr\Big(R(t) \le \hat{R}^{\text{upper}}_{1-\delta}(t)\Big) \ge 1 - \delta. \tag{9}$$

To ensure the statistical validity of this upper bound, we next present an auxiliary lemma (Clopper and Pearson 1934).

Lemma 1 (Clopper–Pearson Confidence Interval for a Binomial Proportion) *Let $W \sim$ Binomial(m, R), and we observe $w \in \{0, 1, \ldots, m\}$ failures. For a significance level $\delta \in (0, 1)$, the Clopper–Pearson two-sided confidence interval for the true (system) failure rate R is $\Big[R^{lower}_{\delta/2},\ R^{upper}_{1-\delta/2}\Big]$, with $\Pr\Big(R \in \Big[R^{lower}_{\delta/2}, R^{upper}_{1-\delta/2}\Big]\Big) \ge 1-\delta$. The two endpoints are defined via the inverse CFD (or quantile) of a Beta distribution: $R^{lower}_{\delta/2} = BetaInv\big(\frac{\delta}{2};\ w,\ m-w+1\big)$, $R^{upper}_{1-\delta/2} = BetaInv\big(1-\frac{\delta}{2};\ w+1,\ m-w\big)$. $BetaInv(p; a, b)$ is the p-th quantile from a beta distribution Beta(a, b) with shape parameters a and b (Wadsworth, Bryan, and Eringen 1961).*

To constrain the worst-case risk with high confidence, we adopt the one-sided Clopper–Pearson bound, corresponding to the upper endpoint of the two-sided interval in Lemma 1. Specifically, given the number of failures w among m i.i.d. Bernoulli trials with failure rate R, then the one-sided upper bound of R with confidence level $1 - \delta$ is defined as:

$$R^{\text{upper}}_{1-\delta} = \text{BetaInv}\Big(1-\delta;\ w+1,\ m-w\Big). \tag{10}$$

We next present the formulation specific to our setting.

Proposition 2. *For a given threshold t, $\hat{m}_{cal}(t)$ is the number of selected calibration samples and $\hat{w}_{cal}(t)$ is the number of failures. Then, the upper bound for the TCFR $R(t)$ is:*

$$\hat{R}^{\text{upper}}_{1-\delta}(t) = \text{BetaInv}\Big(1-\delta;\ \hat{w}_{\text{cal}}(t)+1,\ \hat{m}_{\text{cal}}(t)-\hat{w}_{\text{cal}}(t)\Big), \tag{11}$$

which satisfies $\Pr\Big(R(t) \le \hat{R}^{upper}_{1-\delta}(t)\Big) \ge 1 - \delta$.

While $\hat{R}^{\text{upper}}_{1-\delta}(t)$ is derived in closed form via the Beta inverse CDF in Eq. (11), it is instructive to reinterpret it from the perspective of Bernoulli model presented in Proposition 1. Specifically, we examine the distribution of the empirical failure rate under the unknown conditional risk $R(t)$. This probabilistic view enables us to formulate $\hat{R}^{\text{upper}}_{1-\delta}(t)$ directly in terms of the cumulative distribution of the estimator.

Under the Bernoulli model, the number of failures among the $\hat{m}_{\text{cal}}(t)$ selected samples follows a Binomial distribution with success probability $R(t)$. We denote this random variable as $W \sim \text{Binomial}(\hat{m}_{\text{cal}}(t), R(t))$, and define the empirical failure rate as its normalized form $\hat{R}(t) = \frac{W}{\hat{m}_{cal}(t)}$. That is, $\hat{R}(t)$ represents the *random* empirical failure rate among QA samples with uncertainty no greater than t, and the observed statistic $\hat{r}_{cal}(t)$ in Eq (6) is its realization on the calibration set. To formalize the statistical confidence constraint, we define the CDF of $\hat{R}(t)$ under the true risk $R(t)$ as:

$$D(r \mid R(t)) = \Pr\Big(\hat{R}(t) \le r \mid R(t)\Big). \tag{12}$$

This CDF characterizes the likelihood that the empirical failure rate is no greater than a specified risk threshold r, under the true distribution governed by $R(t)$. With this setup, we identify the largest value of $R(t)$ such that the ECFR $\hat{r}_{\text{cal}}(t)$ remains a statistically plausible observation with probability at least δ, thereby obtaining the following formulation:

$$\hat{R}^{\text{upper}}_{1-\delta}(t) = \sup\{R(t) \in [0,1] : D(\hat{r}_{\text{cal}}(t) \mid R(t)) \ge \delta\}. \tag{13}$$

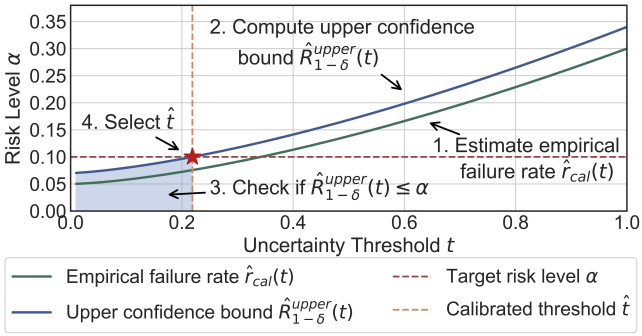


Figure 2: An illustration of the calibration procedure.

Theorem 1. *In the setting of Proposition 1, $\hat{R}^{upper}_{1-\delta}(t)$ defined in Eq.* (13) *unifies both the exact and approximate constructions in Proposition 2 and satisfies the guarantee in Eq.* (9)

Intuitively, since the ECFR $\hat{r}_{cal}(t)$ is already observed on the calibration set, it cannot be a low-probability event under the TCFR $R(t)$. This reasoning leads to the high-probability upper bound $\hat{R}^{\text{upper}}_{1-\delta}(t)$, which ensures that $\hat{r}_{cal}(t)$ remains within the $1-\delta$ quantile of the Binomial distribution. We provide a *formal* proof of Theorem 1 in the appendix.

Threshold Selection. Since the TCFR $R(t)$ is typically non-decreasing in the threshold t, we exploit this monotonicity to identify the largest threshold that satisfies the risk constraint. Formally, we define the calibrated threshold $\hat{t}$ as:

$$\hat{t} = \sup\left\{t : \hat{R}^{\text{upper}}_{1-\delta}(t') \le \alpha \text{ for all } t' \le t\right\}. \tag{14}$$

We select the largest threshold t to retain more QA samples while ensuring that the corresponding upper bound $\hat{R}^{\text{upper}}_{1-\delta}(t)$ is constrained by the risk level α. Notably, this construction obviates the need for conservative multiple hypothesis testing corrections, such as the Bonferroni adjustment, which would otherwise be necessary if the threshold were chosen post hoc without valid uniform guarantees of coverage (Angelopoulos et al. 2021; Quach et al. 2024; Gui, Jin, and Ren 2024). Finally, the calibrated acceptance policy satisfies

$$\Pr\left(R(\hat{t}) \le \hat{R}^{\text{upper}}_{1-\delta}(\hat{t}) \le \alpha\right) \ge 1-\delta, \tag{15}$$

thus ensuring that, with probability at least $1-\delta$, the TCFR of the selected candidate answers (i.e., FDR) remains below the user-specified risk level α under significance level δ.

Implementation Practicality

Computation of the Upper Confidence Bound. In practice, computing $\hat{R}^{\text{upper}}_{1-\delta}(t)$ reduces to solving a one-dimensional root-finding problem: identifying the largest $R(t) \in [0,1]$ such that $\Pr\left(\text{Binomial}\left(\hat{m}_{\text{cal}}(t), R(t)\right) \le \hat{w}_{\text{cal}}(t)\right) \ge \delta$. Given the monotonicity of the Binomial CDF in $R(t)$, the solution can be efficiently obtained via binary search over the interval $[\hat{r}_{\text{cal}}(t), 1]$. This procedure avoids asymptotic approximations and variance estimation, offering robustness and numerical stability in deployment-critical settings.

Calibration of the Threshold. As shown in Figure 2, we determine the largest threshold t (i.e., $\hat{t}$) such that the risk constraint $\hat{R}^{\text{upper}}_{1-\delta}(t) \le \alpha$ is satisfied. This selection leverages the empirical near-monotonicity of the TCFR $R(t)$, allowing us to retain as many samples as possible while ensuring that the final selection policy adheres to the desired risk guarantee.

Scalable Construction of the Confidence Bound. To improve scalability, we investigate approximate yet statistically valid alternatives. As a representative approach, we adopt a concentration-based upper bound derived from Hoeffding's inequality (Hoeffding 1994; Jo 2021; Bates et al. 2021; Angelopoulos et al. 2021), which provides a distribution-free upper confidence bound on the TCFR $R(t)$, using only the empirical failure rate $\hat{r}_{\text{cal}}(t)$ and the number of selected samples $\hat{m}_{\text{cal}}(t)$. We next introduce a supporting lemma.

Lemma 2 (Hoeffding's Inequality for Bernoulli Random Variables) *Let $W_1, \ldots, W_m \in \{0,1\}$ be i.i.d. Bernoulli random variables with common mean $\mu = \mathbb{E}[W_i] \in [0,1]$. Then for any $t \ge 0$, the following inequalities hold:*

$$\Pr\left(\left|\frac{1}{m}\sum_{i=1}^{m} W_i - \mu\right| \ge t\right) \le 2\exp\left(-2mt^2\right). \tag{16}$$

Equivalently, with probability at least $1-\delta$, where $\delta = \exp\left(-2mt^2\right)$, the population mean satisfies

$$\mu \le \frac{1}{m}\sum_{i=1}^{m} W_i + \sqrt{\frac{1}{2m}\log\frac{1}{\delta}}. \tag{17}$$

Let $m = \hat{m}_{\text{cal}}(t)$ and $\hat{r}_{\text{cal}}(t) = \frac{1}{\hat{m}_{\text{cal}}(t)}\sum_{i=1}^{\hat{m}_{\text{cal}}(t)} W_i$. By applying Lemma 2, we obtain a distribution-free, closed-form upper confidence bound on the TCFR $R(t)$ as:

$$\hat{R}^{\text{upper-H}}_{1-\delta}(t) = \hat{r}_{\text{cal}}(t) + \sqrt{\frac{1}{2\hat{m}_{\text{cal}}(t)}\log\frac{1}{\delta}}. \tag{18}$$

We then obtain the Hoeffding-style risk-bounded threshold:

$$\hat{t}^{\text{H}} = \sup\left\{t : \hat{R}^{\text{upper-H}}_{1-\delta}(t') \le \alpha \text{ for all } t' \le t\right\}. \tag{19}$$

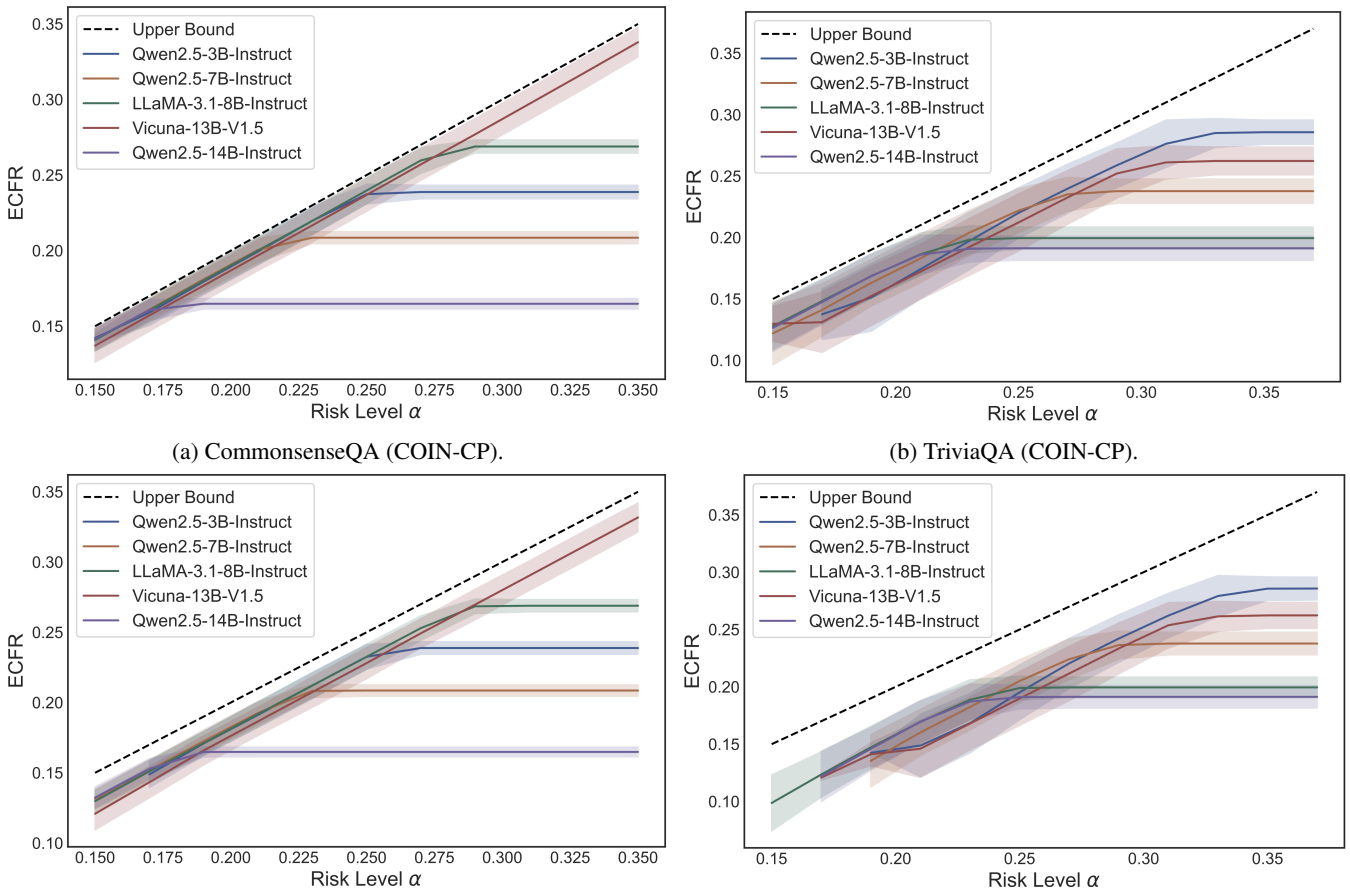
This constitutes a statistically sound yet lightweight substitute for the Clopper–Pearson-based calibration while ensuring user-specified FDR control. A formal proof of Lemma 2 is available in the appendix. Together, these constructions yield the complete and computationally efficient procedure for threshold selection under risk constraints. In the following sections, we refer to the two variants of the COIN framework as COIN-CP and COIN-HFD, corresponding to the Clopper–Pearson and Hoeffding-based bounds, respectively.

Experiments

Experimental Settings

Datasets and Models. We adopt the CommonsenseQA (Talmor et al. 2019) and TriviaQA (Joshi et al. 2017) datasets with five LLMs, and the multimodal MMVet (Yu et al. 2024) dataset utilizing five LVLMs, following (Zhang, Zhang, and Zheng 2024). More details are provided in the appendix.

UQ Methods. COIN accommodates UQ methods under both white-box and black-box settings. By default, we adopt predictive entropy (*PE*) (Kadavath et al. 2022) for closed-ended QA, and semantic entropy (*SE*) (Kuhn, Gal, and Farquhar



(a) CommonsenseQA (COIN-CP).

(b) TriviaQA (COIN-CP).

(c) CommonsenseQA (COIN-HFD).

(d) TriviaQA (COIN-HFD).

Figure 3: Results of the test-time ECFR (i.e., FDR), implementing COIN-CP and COIN-HFD, on both CommonsenseQA and TriviaQA datasets utilizing five LLMs of sizes ranging from 3B to 14B.

2023; Farquhar et al. 2024; Zhang, Zhang, and Zheng 2024) for open-domain QA. In addition, we consider *Ecc*, *Deg*, and *Eigv*, introduced in (Lin, Trivedi, and Sun 2024). More implementation details are provided in the appendix.

Correctness Metrics. Following (Wang et al. 2025b), we use sentence similarity to evaluate correctness in open-domain QA with a threshold of 0.7. We also consider entailment and LLM judgment. More details are provided in the appendix.

Evaluation Metrics. We examine the ECFR on the test set to evaluate the statistical validity of COIN. Following CA (Gui, Jin, and Ren 2024), we also measure power, the proportion of admissible answers in the test set that are retained. We aim to maximize the retention of correct answers under a desired risk level, irrespective of the model's overall accuracy.

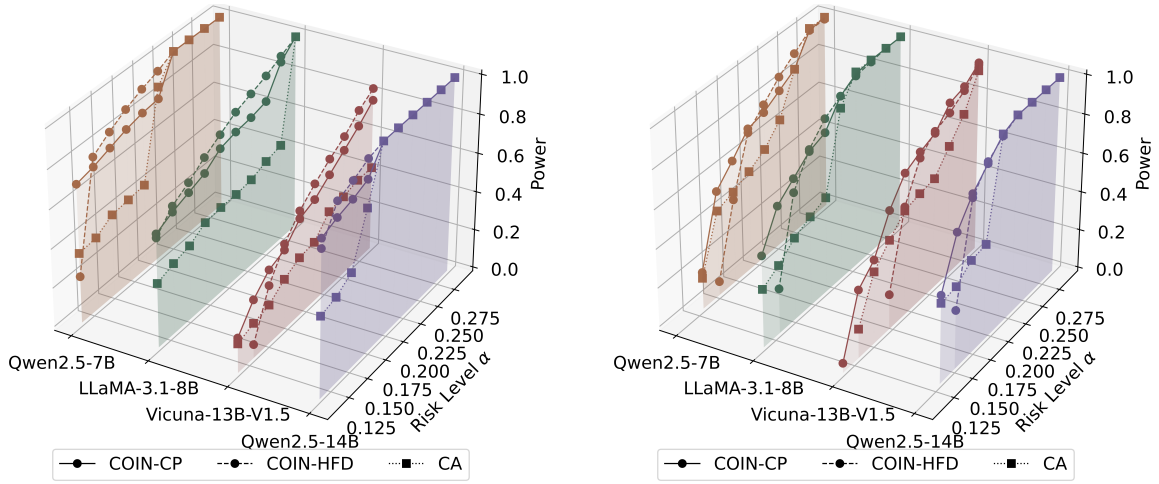
Hyperparameters. Following (Duan et al. 2024; Wang et al. 2025b), we employ beam search (`num_beams=5`) to obtain the most likely generation as the model output. For closed-ended QA, we develop prompts to prompts to elicit the most probable option (e.g., A or B), and set the maximum generation length to 1. For open-domain QA, the generation length is fixed to 32. We set the significance level δ to 0.05. In addition, we fix the calibration-test split ratio to 0.5 by default.

Empirical Evaluations

Statistical Rigor. We evaluate COIN-CP and COIN-HFD on the CommonsenseQA and TriviaQA datasets. ECFR results across five LLMs are presented in Figure 3. *Each solid line denotes the mean ECFR over 100 trials, and the shaded areas indicate $+/-$ the standard deviation.* Both variants rigorously constrain the test-time ECFR under desired risk levels. Given that the Clopper–Pearson method offers an exact, non-asymptotic confidence bound for binomial proportions, while the Hoeffding-based bound is generally looser (Bates et al. 2021), COIN-CP leads to ECFRs tightly concentrated near the target upper bound, whereas COIN-HFD tends to be more conservative, remaining further below the risk level.

We also substantiate the generality, robustness, and efficiency of COIN across various scenarios. Due to space constraints, we provide (1) additional ECFR results on the multimodal MMVet dataset, as well as (2) evaluations incorporating alternative UQ techniques under both white-box and black-box settings, (3) different correctness criteria, and (4) varied calibration-to-test splits in the appendix.

Comparison of Power. While maintaining rigorous risk control, COIN aims to retain as many correct answers as possi-



(a) CommonsenseQA.

(b) TriviaQA.

Figure 4: Comparison of the power metric on both textual QA datasets across 4 LLMs. COIN-CP consistently outperforms CA.

ble. As illustrated in Figure 4, COIN-CP consistently outperforms CA across a range of user-defined risk levels. For instance, in the CommonsenseQA task with Qwen-2.5-14B-Instruct, COIN-CP achieves a power of 0.85 at a low risk level of 0.15, exceeding CA by 0.36, and reaches full power (1.0) at a risk level of just 0.19. On TriviaQA, it outperforms CA by up to 0.4 at the same risk level of 0.19. When using the Hoeffding-style upper bound, COIN-HFD generally surpasses COIN-CP on CommonsenseQA. For instance, with Vicuna-13B-V1.5, COIN-HFD achieves a power of 0.92 at a risk level of 0.25, 0.13 higher than COIN-CP and 0.44 higher than CA. These results highlight that optimizing the upper bound construction in the second stage can substantially improve power without violating risk constraints.

Note that all reported power results are averaged over 100 trials. For each trial, power is counted only if the test-time ECFR falls below the target risk level; otherwise, it is set to zero. The appendix further explores power improvements when more discriminative UQ approaches are adopted in the first stage. These results highlight the modularity and extensibility of the COIN framework: each stage can be independently optimized without compromising theoretical guarantees. Such flexibility allows COIN to accommodate diverse application needs and computational budgets, enabling tailored trade-offs between statistical rigor and practical utility.

Sensitivity Analysis

Given that COIN calibrates thresholds using a held-out calibration set, we investigate its ability to maintain risk control at test time under limited calibration data. As presented in Table 1, COIN consistently satisfies the FDR constraint across a range of target risk levels even with a calibration-to-test split as low as 1:9, highlighting its robustness and predictive efficiency. Furthermore, we demonstrate that COIN remains effective under black-box settings with reduced sampling sizes. We provide additional results to further analyze the sensitivity of COIN to varying calibration-to-test ratios and sampling sizes in the appendix.

Ratio	**0.15**	**0.17**	**0.19**	**0.21**	**0.23**
CommonsenseQA (closed-ended)					
1 : 1	0.1416	0.1607	0.1648	0.1648	0.1648
3 : 7	0.1394	0.1583	0.1648	0.1583	0.1583
1 : 9	0.1305	0.1496	0.1621	0.1643	0.1644
TriviaQA (open-ended)					
1 : 1	0.1264	0.1473	0.1687	0.1859	0.1909
3 : 7	0.1192	0.1388	0.1589	0.1771	0.1871
1 : 9	0.1015	0.1285	0.1516	0.1608	0.1695

Table 1: ECFR Results across various calibration-test split ratios, employing the Qwen2.5-14B-Instruct model.

Conclusion

We have presented COIN, a principled framework that calibrates statistically rigorous thresholds to control FDR when performing selective prediction. COIN consistently achieves higher power than CA in retaining admissible answers. Its modularity enables flexible adaptation to diverse tasks and deployment constraints: ❶ *Stage I:* COIN flexibly integrates either white-box or black-box UQ methods to yield accurate error estimations on the calibration set. ❷ *Stage II:* COIN supports multiple risk control strategies, including the exact Clopper–Pearson interval and scalable alternatives based on concentration inequalities such as Hoeffding's inequality. ❸ *Stage III:* COIN selects the largest threshold to maximize sample retention while satisfying the desired risk constraint. COIN remains valid and efficient even under limited calibration data and black-box settings with small sampling sizes. We envision COIN as a general-purpose framework to advance trustworthy, uncertainty-aware decision-making in foundation models across diverse downstream tasks.

Acknowledgments

Zhiyuan Wang, Xiaofeng Zhu and Xiaoshuang Shi were supported by the National Key Research & Development Program of China under Grant (No. 2022YFA1004100), and the National Natural Science Foundation of China (No.62276052).

References

Abbasli, T.; Toyoda, K.; Wang, Y.; Witt, L.; Ali, M. A.; Miao, Y.; Li, D.; and Wei, Q. 2025. Comparing Uncertainty Measurement and Mitigation Methods for Large Language Models: A Systematic Review. *arXiv preprint arXiv:2504.18346*.

Angelopoulos, A. N.; Barber, R. F.; and Bates, S. 2024. Theoretical foundations of conformal prediction. *arXiv preprint arXiv:2411.11824*.

Angelopoulos, A. N.; and Bates, S. 2021. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*.

Angelopoulos, A. N.; Bates, S.; Candès, E. J.; Jordan, M. I.; and Lei, L. 2021. Learn then test: Calibrating predictive algorithms to achieve risk control. *arXiv preprint arXiv:2110.01052*.

Bates, S.; Angelopoulos, A.; Lei, L.; Malik, J.; and Jordan, M. 2021. Distribution-free, risk-controlling prediction sets. *Journal of the ACM (JACM)*.

Benjamini, Y.; and Hochberg, Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*.

Bentkus, V. 2004. On Hoeffding's Inequalities. *Annals of Probability*, 1650–1673.

Campos, M.; Farinhas, A.; Zerva, C.; Figueiredo, M. A. T.; and Martins, A. F. T. 2024. Conformal Prediction for Natural Language Processing: A Survey. *Transactions of the Association for Computational Linguistics*.

Chen, J.; Yoon, J.; Ebrahimi, S.; Arik, S.; Pfister, T.; and Jha, S. 2023. Adaptation with Self-Evaluation to Improve Selective Prediction in LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2023*.

Cherian, J.; Gibbs, I.; and Candes, E. 2024. Large language model validity via enhanced conformal prediction methods. *Advances in Neural Information Processing Systems*.

Clopper, C. J.; and Pearson, E. S. 1934. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*.

Cresswell, J. C.; Kumar, B.; Sui, Y.; and Belbahri, M. 2025. Conformal Prediction Sets Can Cause Disparate Impact. In *The Thirteenth International Conference on Learning Representations*.

Duan, J.; Cheng, H.; Wang, S.; Zavalny, A.; Wang, C.; Xu, R.; Kailkhura, B.; and Xu, K. 2024. 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*.

Duan, J.; Kong, F.; Cheng, H.; Diffenderfer, J.; Kailkhura, B.; Sun, L.; Zhu, X.; Shi, X.; and Xu, K. 2025. Truthprint: Mitigating lvlm object hallucination via latent truthful-guided pre-intervention. *arXiv preprint arXiv:2503.10602*.

Farquhar, S.; Kossen, J.; Kuhn, L.; and Gal, Y. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*.

Gui, Y.; Jin, Y.; and Ren, Z. 2024. Conformal Alignment: Knowing When to Trust Foundation Models with Guarantees. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Hoeffding, W. 1994. Probability inequalities for sums of bounded random variables. *The collected works of Wassily Hoeffding*.

Huang, L.; Lala, S.; and Jha, N. K. 2024. CONFINE: Conformal Prediction for Interpretable Neural Networks. *arXiv preprint arXiv:2406.00539*.

Huang, L.; Yu, W.; Ma, W.; Zhong, W.; Feng, Z.; Wang, H.; Chen, Q.; Peng, W.; Feng, X.; Qin, B.; et al. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*.

Jaech, A.; Kalai, A.; Lerer, A.; Richardson, A.; El-Kishky, A.; Low, A.; Helyar, A.; Madry, A.; Beutel, A.; Carney, A.; et al. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Jin, Y.; and Candès, E. J. 2023. Selection by prediction with conformal p-values. *Journal of Machine Learning Research*.

Jin, Y.; and Ren, Z. 2024. Confidence on the focal: Conformal prediction with selection-conditional coverage. *arXiv preprint arXiv:2403.03868*.

Jo, T. 2021. Machine learning foundations. *Supervised, Unsupervised, and Advanced Learning. Cham: Springer International Publishing*.

Joshi, M.; Choi, E.; Weld, D.; and Zettlemoyer, L. 2017. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*.

Kadavath, S.; Conerly, T.; Askell, A.; Henighan, T.; Drain, D.; Perez, E.; Schiefer, N.; Hatfield-Dodds, Z.; DasSarma, N.; Tran-Johnson, E.; et al. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Kaur, R.; Samplawski, C.; Cobb, A. D.; Roy, A.; Matejek, B.; Acharya, M.; Elenius, D.; Berenbeim, A. M.; Pavlik, J. A.; Bastian, N. D.; et al. 2024. Addressing Uncertainty in LLMs to Enhance Reliability in Generative AI. In *Neurips Safe Generative AI Workshop 2024*.

Kuhn, L.; Gal, Y.; and Farquhar, S. 2023. Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation. In *The Eleventh International Conference on Learning Representations*.

Kumar, B.; Lu, C.; Gupta, G.; Palepu, A.; Bellamy, D.; Raskar, R.; and Beam, A. 2023. Conformal prediction with large language models for multi-choice question answering. *arXiv preprint arXiv:2305.18404*.

Lin, Z.; Trivedi, S.; and Sun, J. 2024. Generating with Confidence: Uncertainty Quantification for Black-box Large Language Models. *Transactions on Machine Learning Research*.

Mohri, C.; and Hashimoto, T. 2024. Language Models with Conformal Factuality Guarantees. In *International Conference on Machine Learning*.

Ni, B.; Wang, Y.; Cheng, L.; Blasch, E.; and Derr, T. 2025. Towards trustworthy knowledge graph reasoning: An uncertainty aware perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Papadopoulos, H.; Proedrou, K.; Vovk, V.; and Gammerman, A. 2002. Inductive confidence machines for regression. In *Machine Learning: ECML 2002: 13th European Conference on Machine Learning Helsinki, Finland, August 19–23, 2002 Proceedings 13*.

Penny-Dimri, J. C.; Bachmann, M.; Cooke, W. R.; Mathewlynn, S.; Dockree, S.; Tolladay, J.; Kossen, J.; Li, L.; Gal, Y.; and Jones, G. D. 2025. Reducing Large Language Model Safety Risks in Women's Health using Semantic Entropy. *arXiv preprint arXiv:2503.00269*.

Quach, V.; Fisch, A.; Schuster, T.; Yala, A.; Sohn, J. H.; Jaakkola, T. S.; and Barzilay, R. 2024. Conformal Language Modeling. In *The Twelfth International Conference on Learning Representations*.

Shahrokhi, H.; Roy, D. R.; Yan, Y.; Arnaoudova, V.; and Doppa, J. R. 2025. Conformal Prediction Sets for Deep Generative Models via Reduction to Conformal Regression. *arXiv preprint arXiv:2503.10512*.

Singhal, K.; Tu, T.; Gottweis, J.; Sayres, R.; Wulczyn, E.; Amin, M.; Hou, L.; Clark, K.; Pfohl, S. R.; Cole-Lewis, H.; et al. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*.

Snell, J.; Zollo, T. P.; Deng, Z.; Pitassi, T.; and Zemel, R. 2023. Quantile Risk Control: A Flexible Framework for Bounding the Probability of High-Loss Predictions. In *The Eleventh International Conference on Learning Representations*.

Su, W.; Wang, C.; Ai, Q.; Hu, Y.; Wu, Z.; Zhou, Y.; and Liu, Y. 2024. Unsupervised Real-Time Hallucination Detection based on the Internal States of Large Language Models. In *Findings of the Association for Computational Linguistics: ACL 2024*.

Talmor, A.; Herzig, J.; Lourie, N.; and Berant, J. 2019. 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*.

Tayebati, S.; Kumar, D.; Darabi, N.; Jayasuriya, D.; Krishnan, R.; and Trivedi, A. R. 2025. Learning Conformal Abstention Policies for Adaptive Risk Management in Large Language and Vision-Language Models. *arXiv preprint arXiv:2502.06884*.

Wadsworth, G. P.; Bryan, J. G.; and Eringen, A. C. 1961. Introduction to probability and random variables. *Journal of Applied Mechanics*.

Wang, Q.; Geng, T.; Wang, Z.; Wang, T.; Fu, B.; and Zheng, F. 2025a. Sample then Identify: A General Framework for Risk Control and Assessment in Multimodal Large Language Models. In *The Thirteenth International Conference on Learning Representations*.

Wang, Z.; Duan, J.; Cheng, L.; Zhang, Y.; Wang, Q.; Shi, X.; Xu, K.; Shen, H. T.; and Zhu, X. 2024. ConU: Conformal Uncertainty in Large Language Models with Correctness Coverage Guarantees. In *Findings of the Association for Computational Linguistics: EMNLP 2024*.

Wang, Z.; Duan, J.; Yuan, C.; Chen, Q.; Chen, T.; Zhang, Y.; Wang, R.; Shi, X.; and Xu, K. 2025b. Word-sequence entropy: Towards uncertainty estimation in free-form medical question answering applications and beyond. *Engineering Applications of Artificial Intelligence*.

Wang, Z.; Wang, Q.; Zhang, Y.; Chen, T.; Zhu, X.; Shi, X.; and Xu, K. 2025c. Sconu: Selective conformal uncertainty in large language models. *arXiv preprint arXiv:2504.14154*.

Yadkori, Y. A.; Kuzborskij, I.; Stutz, D.; György, A.; Fisch, A.; Doucet, A.; Beloshapka, I.; Weng, W.-H.; Yang, Y.-Y.; Szepesvári, C.; et al. 2024. Mitigating llm hallucinations via conformal abstention. *arXiv preprint arXiv:2405.01563*.

Yang, R.; Zhang, C.; Zhang, Z.; Huang, X.; Yu, D.; Collier, N.; and Yang, D. 2025. UNCLE: Uncertainty Expressions in Long-Form Generation. *arXiv preprint arXiv:2505.16922*.

Ye, F.; Yang, M.; Pang, J.; Wang, L.; Wong, D.; Yilmaz, E.; Shi, S.; and Tu, Z. 2024. Benchmarking llms via uncertainty quantification. *Advances in Neural Information Processing Systems*.

Yu, W.; Yang, Z.; Li, L.; Wang, J.; Lin, K.; Liu, Z.; Wang, X.; and Wang, L. 2024. MM-Vet: Evaluating Large Multimodal Models for Integrated Capabilities. In *Forty-first International Conference on Machine Learning*.

Zhang, R.; Zhang, H.; and Zheng, Z. 2024. VL-Uncertainty: Detecting Hallucination in Large Vision-Language Model via Uncertainty Estimation. *arXiv preprint arXiv:2411.11919*.

Zheng, Y.; Gan, W.; Chen, Z.; Qi, Z.; Liang, Q.; and Yu, P. S. 2025. Large language models for medicine: a survey. *International Journal of Machine Learning and Cybernetics*.

Zollo, T. P.; Morrill, T.; Deng, Z.; Snell, J.; Pitassi, T.; and Zemel, R. 2024. Prompt Risk Control: A Rigorous Framework for Responsible Deployment of Large Language Models. In *The Twelfth International Conference on Learning Representations*.