CAP: Conformalized Abstention Policies for Context-Adaptive Risk Management for LLMs and VLMs

Sina Tayebati STAYEB3@UIC.EDU
Divake Kumar DKUMAR33@UIC.EDU
Nastaran Darabi NDARAB2@UIC.EDU
Dinithi Jayasuriya DKASTH2@UIC.EDU
Theja Tulabandhula THEJA@UIC.EDU
University of Illinois Chicago, Illinois, United States
Ranganath Krishnan* RANGANATH.KRISHNAN@CAPITALONE.COM
AI Labs, Capital One, Texas, United States
Amit Ranjan Trivedi AMITRT@UIC.EDU
University of Illinois Chicago, Illinois, United States

Editors: Hung-yi Lee and Tongliang Liu

Abstract

Large Language and Vision-Language Models (LLMs/VLMs) are increasingly deployed in high-stakes domains where predictive failures can be costly. Conformal Prediction (CP) offers distribution-free uncertainty quantification with finite-sample coverage guarantees, but its reliance on a globally fixed risk level enforces a uniform trade-off between coverage and informativeness, misaligned with the instance-specific uncertainty patterns of modern foundation models. We propose the framework of *Conformalized Abstention Policy* (CAP), a novel framework that integrates CP with deep Reinforcement Learning (RL) to learn per-instance abstention policies. CAP trains a utility-driven policy to dynamically select the conformal risk level for each input, balancing point prediction, set prediction, and full abstention based on downstream utility. We specifically introduce *Policy-Calibrated Coverage*, a theoretical guarantee ensuring that the empirical coverage of the learned policy reliably estimates its true expected performance. Extensive experiments show that CAP maintains the 90% target coverage while substantially outperforming static CP baselines: improving hallucination detection AUROC by up to 22.2%, uncertainty-guided selective generation AUARC by 21.2%, and reducing calibration error by over 70%. CAP also extends to free-form generation by managing the trade-off between a detailed and factual response on a per-instance basis by learning an optimal risk level for sub-claim retention.

Keywords: Conformal prediction; LLM/VLM; Risk Management

1. Introduction

Large Language and Vision-Language Models (LLMs/VLMs) are increasingly deployed in high-stakes domains, from medical diagnostics to autonomous navigation, where predictive failures can have severe consequences. Despite their advanced capabilities, these models remain prone to complex and semantically nuanced failures, such as factual hallucinations, harmful biases, and unpredictable reasoning errors, which are difficult to anticipate and

* Work done while at Intel Labs

mitigate Abdali et al. (2024). A critical requirement for their safe deployment is the ability to reliably quantify uncertainty and abstain from making predictions when confidence is low.

Conformal Prediction (CP) has emerged as a powerful framework for uncertainty quantification, offering distribution-free, finite-sample coverage guarantees Vovk et al. (2005). By constructing prediction sets guaranteed to contain the true label with a user-specified probability (e.g., 90%), CP provides a principled mechanism for risk management. However, standard CP applies a globally fixed risk level α, imposing a uniform trade-off between coverage and informativeness (i.e., prediction set size) across all inputs. This one-size-fits-all approach is fundamentally misaligned with modern foundation models, where the nature and severity of potential failures vary dramatically across inputs. For example, a model may be highly certain on simple factual queries but dangerously overconfident on complex multi-step reasoning. Static risk control fails to capture such *instance-specific uncertainty*.

This presents a central challenge: *how can we move beyond static, global risk levels toward adaptive, per-instance risk assessment that responds to the uncertainty inherent in each input?* Existing adaptive methods fall short. Heuristic approaches lack formal guarantees, while methods such as Adaptive Prediction Sets (APS) Romano et al. (2020) and Least Ambiguous Classifiers (LAC) Sadinle et al. (2019) optimize set size but do not learn a data-driven abstention policy for a specific utility function. Importantly, the essential goal is not only to construct prediction sets, but to decide when to abstain, when to return a set, and when to make a confident point prediction, thereby optimizing downstream utility.

To address this gap, we introduce the *Conformalized Abstention Policy* (CAP), a framework that frames adaptive risk management as a learned, utility-maximizing decision process. CAP integrates the formal guarantees of CP with the adaptive power of Reinforcement Learning (RL). By learning an instance-conditional policy, CAP dynamically selects the risk level for each input, trading off coverage, precision, and abstention to maximize a defined utility function. Our contributions are:

1. Framework for Learned Conformal Abstention. We propose the first principled integration of deep RL with CP for adaptive risk control in LLMs/VLMs. We formulate the selection of the conformal risk level α as a learned, instance-conditional policy that maps inputs to optimal risk levels to control abstention.

2. Utility-Driven Policy Optimization. We introduce a dual-threshold conformal mechanism, controlled by an RL agent, that learns a policy to choose among confident prediction, partial abstention (returning a set), or full abstention. This policy is optimized to maximize a flexible utility function that balances the costs of different errors, bridging statistical guarantees with application needs.

3. Theoretical and Empirical Validation. We provide a formal guarantee of *Policy-Calibrated Coverage*, showing that our learned policy's empirical coverage is a reliable estimate of its true expected performance. We validate CAP on hallucination detection and selective generation benchmarks, where it substantially improves performance: boosting hallucination detection by up to 22%, improving selective generation by over 20%, and reducing calibration error by 70–85% versus static baselines, all while maintaining guarantees. We also show CAP's efficacy in free-form factuality checking (FActScore, NQ, MATH), where it retains 8–13% more factual sub-claims than the best prior conformal baseline while holding factuality at 90%.

2. Related Works

2.1. Uncertainty Quantification in Foundation Models

The deployment of Foundation Models such as LLMs and VLMs in high-stakes applications has made Uncertainty Quantification (UQ) a critical area of research Gawlikowski et al. (2023). Classical UQ paradigms, however, face significant scalability challenges. Bayesian neural networks, while offering a principled measure of uncertainty, are difficult to scale, as true posterior inference over billions of parameters is computationally intractable. Practical approximations such as Monte Carlo dropout Gal and Ghahramani (2016) and deep ensembles Lakshminarayanan et al. (2017) require multiple forward passes at inference time, which is prohibitively expensive for foundation-scale models. As a result, research has shifted toward more efficient UQ methods, which can be broadly categorized into intrinsic and extrinsic.

Intrinsic methods leverage the model's own outputs, such as the entropy of the softmax distribution or verbalized confidence, where the model is prompted to express its certainty Huang et al. (2023). A notable recent example is semantic entropy, which measures diversity in the semantic space of multiple generated samples to detect hallucinations Farquhar et al. (2024). Extrinsic methods include post-hoc calibration techniques, such as temperature scaling and Platt scaling, to align model probabilities with empirical correctness Kadavath et al. (2022). Recent approaches also explore prompting-based calibration Zhao et al. (2021) and selective decoding to trade off coverage for correctness in language generation Lee et al. (2024). While these techniques are effective for ranking predictions by confidence, they lack the formal, distribution-free statistical guarantees needed for reliable risk management. This limitation motivates our use of CP or actionable risk control.

2.2. Conformal Prediction in LLMs/VLMs

CP has emerged as a powerful framework for providing rigorous, model-agnostic uncertainty guarantees Angelopoulos and Bates (2021). By calibrating model outputs on a hold-out set, it constructs prediction sets that are guaranteed to contain the true label with a user-specified probability. The key challenge in applying CP to LLMs and VLMs lies in adapting a framework originally designed for structured classification and regression to the inherently unstructured and variable-length nature of text and multimodal generation.

A major line of work addresses this by designing novel non-conformity scores for sequences, enabling *conformal language modeling* Quach et al. (2023). Other related approaches extend CP to handle generative tasks, including selective decoding with conformal guarantees Lee et al. (2024), and have begun to explore CP-based methods for multimodal models as well Tumu et al. (2024). Other research focuses on leveraging CP for selective classification and hallucination mitigation. For example, Cherian et al. (2024) improves the underlying confidence estimates used by CP to enhance abstention performance. Closer to our work, Yadkori et al. (2024) uses CP to trigger abstention based on the semantic consistency of multiple sampled responses, but employs a single, globally-fixed risk threshold for abstention.

While these approaches have advanced the field, they either focus on engineering improved non-conformity scores or apply a static risk-control policy. In contrast, our work introduces a different direction: we propose learning a *dynamic, instance-adaptive policy* that selects the operational risk level (α) on a per-instance basis. By formulating this as an RL problem, CAP

learns an explicit, context-aware trade-off between coverage, accuracy, and informativeness, moving beyond static guarantees to achieve adaptive risk control.

2.3. Selective Classification and Learning with Rejection

Selective classification, or "learning with a reject option," enables models to abstain when uncertain, thereby enabling a principle foundational to trustworthy AI. Chow's seminal work Chow (1970) established that a Bayes-optimal classifier should reject when the maximum posterior falls below a threshold. The goal is to optimize a utility function that balances prediction errors against the cost of abstention Hendrickx et al. (2024). This binary predict-or-reject setup has been extended to partial abstention, where models output a subset of plausible labels Pugnana et al. (2024), aligning naturally with the set-valued outputs of CP, which provides rigorous coverage guarantees.

Recent efforts have integrated selective classification with CP's statistical validity. For instance, Sadinle et al. Sadinle et al. (2019) introduced the LAC, which minimizes set size while ensuring coverage. However, such approaches still assume globally fixed risk-reward trade-offs. Additional work on selective prediction and abstention includes SelectiveNet Geifman and El-Yaniv (2019), which jointly trains for prediction and rejection, and Bayesian ensemble methods for confidence-based abstention Mukhoti et al. (2023). In the context of LLMs, recent efforts explore abstention via semantic consistency Yadkori et al. (2024) and explicit rejection prompts Varshney et al. (2022), though without formal guarantees.

CAP offers a principled extension. Instead of static thresholds tied to a global utility function, CAP uses RL to learn a flexible, input-dependent abstention policy. By mapping each input to optimal conformal parameters (α, β), CAP dynamically chooses between confident prediction, set-valued output, or full abstention. This allows CAP to maximize an empirical reward that reflects instance-level trade-offs, enabling adaptive and effective risk management in foundation models.

3. Conformalized Abstention Policy (CAP)

We propose CAP, a framework that combines the statistical guarantees of CP with the adaptive decision-making capabilities of RL. In standard CP, the risk level α is globally fixed, imposing a uniform trade-off between coverage and informativeness across all inputs. This is suboptimal for LLMs/VLMs, whose failure modes are complex and semantically nuanced. CAP introduces an *instance-conditional policy*, $\pi_\theta(\alpha|\mathbf{x})$, that dynamically selects the risk level per input, adapting to varying uncertainty and risk. This reframes risk selection as a learned utility maximization problem, allowing the policy to explicitly model input-dependent failure modes. Unlike prior conformal abstention methods based on static thresholds or heuristic score calibration, CAP provides the first principled integration of RL-driven abstention with CP's statistical guarantees, specifically for LLM/VLM settings.

3.1. Leveraging Conformal Prediction for Adaptive Risk Control

Standard inductive CP provides finite-sample, distribution-free coverage guarantees. Given a calibration set $\mathcal{D}_{\text{cal}} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ and a non-conformity score $s(\mathbf{x}, y)$, scores $s_i = s(\mathbf{x}_i, y_i)$ are computed. For a desired risk level α, the threshold $\hat{q}$ is the $\frac{\lceil (n+1)(1-\alpha) \rceil}{n}$-th quantile

of calibration scores. The resulting prediction set $\Gamma(\mathbf{x}_t) = \{y \in \mathcal{Y} : s(\mathbf{x}_t, y) \leq \hat{q}\}$ satisfies $\mathbb{P}(y_t \in \Gamma(\mathbf{x}_t)) \geq 1 - \alpha$. We leverage this framework by casting adaptive risk management as a decision-theoretic problem for CAP. For any input $\mathbf{x}$, the model must select an action: provide a single confident prediction, return a set of plausible options (partial abstention), or abstain entirely. The optimal choice depends on the utility of each outcome.

While prior works have explored static or globally-optimized abstention thresholds, these approaches cannot adapt to the highly variable uncertainty landscape of LLM/VLM outputs. CAP addresses this by introducing a learnable, instance-conditional abstention mechanism driven by RL, which jointly optimizes predictive utility and formal risk control. We formalize this via a utility function $U(y, \mathcal{O})$, where $\mathcal{O}$ is the model's output given the true label y. The output $\mathcal{O}$ is a tuple $(\hat{y}, \Gamma, a)$, where $\hat{y} \in \mathcal{Y}$ is a point prediction, $\Gamma \subseteq \mathcal{Y}$ is a prediction set, and $a \in \{\text{predict, set, abstain}\}$ is the action. The utility is defined as:

$$U(y, \mathcal{O}) = -\Big(c_{\text{err}}\mathbb{I}(\hat{y} \neq y \wedge a = \text{predict}) + c_{\text{set}}(|\Gamma|)\mathbb{I}(a = \text{set}) + c_{\text{abs}}\mathbb{I}(a = \text{abstain})\Big), \quad (1)$$

where c_{err} is the cost of a misprediction, $c_{\text{set}}(|\Gamma|)$ penalizes large prediction sets (lack of informativeness), and c_{abs} is the cost of abstention. Thereby, the goal of CAP is to learn a policy $\pi : \mathcal{X} \to \{\text{predict, set, abstain}\}$ that maximizes expected utility:

$$\pi^* = \arg\max_{\pi} \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[U(y, \pi(\mathbf{x}))]. \quad (2)$$

3.2. Policy-Driven Risk Control via Conformal RL

Directly optimizing Equation 2 is intractable, as it requires knowledge of y at test time. CAP addresses this by using a dual-threshold conformal mechanism to define the action space and RL to learn a policy that approximates utility maximization in a data-driven manner. The policy π_θ learns an instance-conditional risk vector $(\alpha, \beta) \in [0, 1]^2$ for each input $\mathbf{x}$. These parameters configure a dual-threshold conformal system that partitions model uncertainty, as measured by the non-conformity score $s(\mathbf{x})$, into three actions: confident prediction, partial abstention, or full abstention. The thresholds $\hat{q}_{\text{predict}}(\alpha)$ and $\hat{q}_{\text{abstain}}(\beta)$ are computed from calibration scores $\{s_i\}_{i=1}^n$ derived from a calibration set $\mathcal{D}_{\text{cal}}$.

This setup enables utility maximization via policy gradient RL. For each input $\mathbf{x}$ (state), the agent samples an action (α, β) from its policy, configures the conformal system accordingly, and receives a reward based on performance on a batch of data. The reward serves as an empirical proxy for the utility function U. It is defined as the negative of the following cost:

$$C(\alpha, \beta) = (1 - \text{acc}) + \lambda_1\text{avgSet} + \lambda_2\text{abstention} - \lambda_3\text{coverage} - \lambda_4\text{div}, \quad (3)$$

with $R(\alpha, \beta) = -C(\alpha, \beta)$. Each term reflects the utilities of our framework: $(1 - \text{acc})$ estimates c_{err}, while avgSet and abstention approximate c_{set} and c_{abs}, respectively. The reward includes a coverage bonus and an information-theoretic regularizer (Shannon entropy, div) to promote exploration. We note that the reward is non-stationary during training, as statistics such as acc and avgSet depend on the evolving policy π_θ. However, we find that a standard policy gradient approach with a sufficiently small learning rate yields stable training and converges effectively to a policy that maximizes expected utility.

3.3. Policy Optimization via Stochastic Gradient Ascent

The policy parameters θ are optimized to maximize the expected utility defined in Equation 2, which we achieve by maximizing the expected reward $J(\theta) = \mathbb{E}_{\mathbf{x}\sim\mathcal{D},(\alpha,\beta)\sim\pi_\theta(\cdot|\mathbf{x})}[R(\alpha,\beta)]$. Because the reward depends on non-differentiable components (quantile functions and discrete metrics), we employ a model-free policy gradient algorithm. Specifically, we use REINFORCE to update the policy by ascending the gradient of the expected reward, which allows us to optimize θ without requiring differentiability of the reward function. For our setting, REINFORCE provides a Monte Carlo estimate of this gradient. The expectation is approximated by sampling from the policy and the data distribution. For each input $\mathbf{x}_t$, we sample an action $a_t = (\alpha_t, \beta_t) \sim \pi_\theta(\cdot|\mathbf{x}_t)$ and compute the resulting reward $R_t = R(\alpha_t, \beta_t)$. The gradient is then approximated as:

$$\nabla_\theta J(\theta) \approx R(\alpha,\beta)\nabla_\theta \log \pi_\theta(\alpha,\beta|\mathbf{x}). \tag{4}$$

The policy parameters are updated via stochastic gradient ascent on a mini-batch of samples:

$$\theta_{t+1} \leftarrow \theta_t + \eta \cdot \left(\frac{1}{N}\sum_{i=1}^{N} R_i \nabla_\theta \log \pi_\theta(a_i|s_i)\right), \tag{5}$$

where η is the learning rate, N is the batch size, and R_i is the reward obtained from action a_i in state s_i. While the REINFORCE gradient estimator is unbiased, it is known to exhibit high variance. To mitigate this, we incorporate a baseline $b(\mathbf{x})$, subtracting it from the reward: $\nabla_\theta J(\theta) \approx (R(\alpha,\beta) - b(\mathbf{x}))\nabla_\theta \log \pi_\theta(\alpha,\beta \mid \mathbf{x})$. We use the moving average of rewards as the baseline, which significantly stabilizes training. This update couples the conformal thresholds to the RL objective to learn a principled trade-off between predictive certainty and abstention risk. The full procedure is summarized in supplementary material.

3.4. Theoretical Guarantees on the RL-Learned Abstention Policy

A central theoretical question is how CAP's adaptivity interacts with the formal guarantees of CP. Our framework inherits a rigorous *conditional* coverage guarantee directly from CP. For any risk level $\alpha(\mathbf{x})$ selected by the policy for input $\mathbf{x}$, the resulting prediction set $\Gamma(\mathbf{x};\alpha(\mathbf{x}))$ satisfies, under exchangeability:

$$\mathbb{P}\left(Y \in \Gamma(\mathbf{x};\alpha(\mathbf{x})) \,\middle|\, \mathbf{x}\right) \geq 1 - \alpha(\mathbf{x}). \tag{6}$$

This ensures that each decision to produce a prediction set is backed by a valid statistical guarantee for its chosen risk level. We discuss the following constructs towards this:

Conditional Coverage Guarantee. For any input $\mathbf{x}$, the policy-selected risk level $\alpha(\mathbf{x})$ yields a conformal prediction set $\Gamma(\mathbf{x};\alpha(\mathbf{x}))$ such that:

$$\mathbb{P}\left(Y \in \Gamma(\mathbf{x};\alpha(\mathbf{x})) \mid \mathbf{x}\right) \geq 1 - \alpha(\mathbf{x}). \tag{7}$$

However, this conditional result does not fully characterize the aggregate, long-run behavior of an adaptive policy. Practitioners require assurance that observed performance on finite datasets reliably reflects the policy's expected behavior. We address this with two complementary results: a bound on the deviation between target and true expected coverage, and a high-confidence guarantee on empirical coverage estimates.

Proposition 1 (Bound on Expected Coverage Deviation). Let π_θ be a fixed policy, and assume the conformal p-values $p(\mathbf{x}, y)$ are continuously distributed. Then the expected coverage of the policy satisfies:

$$\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathrm{Cov}(\pi_\theta)] \geq 1 - \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}[\alpha(\mathbf{x})] - \mathcal{O}\left(\frac{1}{n}\right), \tag{8}$$

where $n = |\mathcal{D}_{\mathrm{cal}}|$ is the size of the calibration set, and the expectation is over the data.
Proof sketch. For each input $\mathbf{x}$ and chosen threshold $\alpha(\mathbf{x})$, the conformal coverage is given by $P(p(\mathbf{x}, Y) > \alpha(\mathbf{x})) \geq 1 - \alpha(\mathbf{x})$. This inequality holds in finite samples by CP's marginal validity. Averaging over the randomness of the calibration set yields a deviation of order $\mathcal{O}(1/n)$. Taking the expectation over $\mathbf{x} \sim \mathcal{D}$ completes the result.

Proposition 2 (Policy-Calibrated Coverage). We introduce the novel notion of policy-calibrated coverage. Let π_θ be a fixed policy learned by CAP. For a test set $\mathcal{D}_{\mathrm{test}} = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$ of m i.i.d. samples, define the empirical coverage as:

$$\mathrm{Cov}_m(\pi_\theta) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{I}\left(y_i \in \Gamma(\mathbf{x}_i; \alpha(\mathbf{x}_i))\right). \tag{9}$$

Then, for any $\delta > 0$, the following concentration inequality holds:

$$\mathbb{P}\left(|\mathrm{Cov}_m(\pi_\theta) - \mathbb{E}[\mathrm{Cov}(\pi_\theta)]| \leq \sqrt{\frac{\ln(2/\delta)}{2m}}\right) \geq 1 - \delta. \tag{10}$$

This result, a direct application of Hoeffding's inequality, ensures that the empirical coverage observed on any sufficiently large test set is a reliable estimator of the policy's long-run performance. It provides a crucial confidence guarantee: empirical metrics reported in practice are not artifacts of chance but concentrate tightly around their expected value.

Together, Propositions 1 and 2 offer a strong theoretical foundation for CAP: it combines the local optimality and distribution-free guarantees of CP with the global adaptivity of RL. While the theoretical gap between $\mathbb{E}[\mathrm{Cov}(\pi_\theta)]$ and $1 - \mathbb{E}[\alpha(\mathbf{x})]$ remains an open question, our results demonstrate that empirically learned policies can produce reliable, statistically valid predictions in practice. Future work may pursue tighter guarantees using risk-sensitive regularization, variance-reduced policy gradients, or conformal scores tailored to generative LLM outputs, which is currently an open question.

4. Experiments

4.1. Experimental Setup

Datasets: We design our evaluation framework around multiple-choice question-answering (MCQA) to enable standardized and controlled comparison of uncertainty estimates. We use ten benchmark datasets spanning diverse reasoning tasks. For VLMs, we evaluate five datasets reformatted to four- or six-choice MCQA: **MMBench**, **OODCV-VQA** (Digits subset), **ScienceQA**, **SEEDBench**, and **AI2D**. For LLMs, we use five tasks standardized to six-choice format (including "I don't know" as an explicit abstention option): **MMLU**, **CosmosQA**, **HellaSwag**, **HaluDial**, and **HaluSum**. This selection spans knowledge

recall, multi-hop reasoning, and robustness under ambiguity, providing a comprehensive test bed for abstention and uncertainty evaluation.

Models: We evaluate a range of LLMs and VLMs spanning 2.7B to 34B parameters. For VLMs, we report results for the LLaVA-v1.6 series (34B, 13B, 7B). For LLMs, we include Yi-34B and Qwen (7B, 14B). We briefly describe the characteristics of these models and datasets in supplementary material.

Implementation Details: Each dataset is split 50% for the calibration set ($\mathcal{D}_{\text{cal}}$) and 50% for the test set. The policy network f_θ is a lightweight multi-layer perceptron (MLP), trained for 200 epochs with learning rate $\eta = 0.001$. The scaling constant c in the sigmoid functions for action probability calculation is set to 5. The cost function weights are: $\lambda_1 = 0.3$ (set size), $\lambda_2 = 0.3$ (abstention), $\lambda_3 = 0.4$ (coverage), $\lambda_4 = 0.2$ (diversity), with an additional penalty weight $\lambda_5 = 0.5$ to discourage extreme policy outputs. For the tested LLM/VLMs, temperature is set to 0.3, with $top - p = 0.9$.

Evaluation Metrics: We evaluate CAP and baseline methods using the following metrics to assess both predictive performance and the reliability of UQ:

Accuracy: For a test input X_t with true label Y_t, let $C(X_t)$ denote the generated prediction set. If a single prediction $\hat{Y}_t$ is produced (e.g., in confident scenarios under CAP), accuracy is binary: 1 if $\hat{Y}_t = Y_t$, and 0 otherwise. For set predictions, accuracy is computed fractionally, inversely proportional to the size of $C(X_t)$ when $Y_t \in C(X_t)$.

Coverage: Coverage measures the fraction of instances where the true label is included in the model's output—either as a single prediction or within a prediction set. In abstention setups, coverage also accounts for instances where the model abstains from an incorrect explicit prediction. This metric verifies whether the conformal method achieves target coverage.

Set Sizes (SS): Set Size measures the average number of labels in prediction sets, excluding single predictions and abstentions. It reflects model uncertainty: larger sets indicate higher uncertainty, smaller sets indicate greater confidence.

Area Under the Receiver Operating Characteristic (AUROC): AUROC evaluates the model's ability to distinguish correct from incorrect predictions based on confidence scores Davis and Goadrich (2006). A higher AUROC indicates more reliable ranking of predictions, critical for tasks such as hallucination detection.

Area Under the Accuracy-Rejection Curve (AUARC): AUARC measures the area under the accuracy-rejection curve, capturing the trade-off between predictive accuracy and the retained data fraction Krishnan et al. (2024). A higher AUARC indicates a more effective uncertainty estimate, enabling to abstain on difficult (likely predictively inaccurate) instances.

Expected Calibration Error (ECE): ECE quantifies the discrepancy between a model's average prediction confidence and its empirical accuracy Naeini et al. (2015). It is computed as the weighted average of the absolute difference between accuracy and confidence across bins:

$$\text{ECE} = \sum_{b=1}^{n_{\text{bins}}} \frac{|B_b|}{N} \left|\text{acc}(B_b) - \text{conf}(B_b)\right|. \tag{11}$$

A lower ECE indicates that the model's confidence scores are better calibrated and more aligned with the true likelihood of correctness.

Prompting Strategies: To ensure standardized evaluation, we use specific prompting templates for both VLMs and LLMs. For **VLMs**, we adapt the MCQA template from LLaVA Liu et al. (2024). Each prompt includes the input image, the question, and a list of lettered options (e.g., A–F). The model is instructed to return only the letter corresponding to its selected answer (e.g., "A"), yielding a well-defined logit distribution over the fixed choices. We use official model-specific templates to ensure compatibility and omit few-shot examples due to typical single-image input constraints. For **LLMs**, we use the zero-shot *Base Prompt* strategy Zheng et al. (2023), which concatenates the question and all options into a single prompt and instructs the model to provide its answer following the prefix "Answer:". Detailed examples of prompt templates are provided in supplementary material.

4.2. Results and Analysis

4.2.1. Hallucination Detection and Selective Generation

We assess CAP's ability to detect hallucinations in MCQA tasks, using AUROC as the primary metric Farquhar et al. (2024) to reflect the ranking quality of uncertainty signals. Table 1 shows that CAP consistently outperforms two strong baselines—APS Romano et al. (2020) and LAC Sadinle et al. (2019)—across all datasets and model scales.

For VLMs, CAP improves average AUROC by up to **10.17%** over APS and **8.12%** over LAC, with peak gains of **22.19%** on SQA using LLaVA-13B. These gains are especially notable given the formal coverage guarantees of the baselines. CAP's edge stems from its ability to adapt abstention thresholds to instance-specific features, capturing finer-grained semantic uncertainty than global or class-wise methods. Similar trends hold for LLMs, where CAP yields strong gains on HDial, CQA, and HSwg, including for smaller models like Qwen-7B where baseline performance is weaker. This suggests CAP not only better aligns with risk-awareness but also mitigates epistemic uncertainty in lower-capacity models. These results support our central claim: combining CP with input-conditional abstention policies leads to statistically grounded and behaviorally robust improvements.

We also evaluated each method's ability to support selective generation—abstaining from uncertain predictions to improve reliability. Performance is measured using AUARC Farquhar et al. (2024); Krishnan et al. (2024), which quantifies how effectively uncertainty estimates enable graceful degradation—i.e., maintaining high accuracy on retained predictions as abstention increases. Higher AUARC indicates alignment between uncertainty and error.

As shown in Table 1, CAP consistently outperforms APS and LAC across all evaluated VLMs and LLMs. It improves average AUARC by up to **9.43%**, with peak gains of **21.17%** on MMB using LLaVA-13B. These improvements hold across both high-capacity models (e.g., Yi-34B, LLaVA-34B) and lower-capacity ones (e.g., Qwen-7B), demonstrating CAP's robustness across architectures and domains. In these results, CAP yields steeper accuracy-rejection curves and better utility under constraint. By abstaining in a risk-aware, utility-preserving manner, CAP enhances reliability in settings like dialogue, reasoning, and multimodal QA where overgeneration is costly, establishing a new standard for principled uncertainty-guided generation in LLMs and VLMs.

4.2.2. Coverage Guarantee and Set Size Trade-off

A fundamental requirement for any conformal prediction method is to satisfy the target marginal coverage with finite-sample guarantees. As evidenced in Table 2, CAP consistently

Table 1: Comparison of CAP (ours) with Benchmarking LLMs via Uncertainty Quantification Ye et al. (2024), and Uncertainty-Aware Evaluation Kostumov et al. (2024). Results are reported for VLMs and LLMs, evaluated on AUROC (hallucination detection) and AUARC (uncertainty-guided selective generation). Best values are shown in **bold**.

Models	**Method**	**AUROC ↑**						**AUARC ↑**					
VLMs		**MMB**	**OOD**	**SQA**	**SB**	**AI2D**	**Avg.**	**MMB**	**OOD**	**SQA**	**SB**	**AI2D**	**Avg.**
LLaVA-v1.6-34B	APS	0.72	0.70	0.72	0.56	0.84	0.71	0.96	0.91	0.93	0.92	0.93	0.93
	LAC	0.78	0.70	0.80	0.56	0.85	0.74	0.94	0.90	0.91	0.88	0.92	0.91
	VLM-Bench	0.75	0.70	0.76	0.56	0.85	0.73	0.95	0.91	0.92	0.90	0.93	0.92
	Ours	**0.80**	**0.78**	**0.86**	**0.65**	**0.90**	**0.80**	**0.98**	**0.97**	**0.98**	**0.94**	**0.99**	**0.97**
LLaVA-v1.6-13B	APS	0.49	0.59	0.53	0.49	0.78	0.57	0.96	0.88	0.94	0.93	0.91	0.92
	LAC	0.68	0.59	0.60	0.50	0.75	0.63	0.93	0.85	0.89	0.88	0.90	0.89
	VLM-Bench	0.59	0.59	0.57	0.50	0.77	0.60	0.95	0.87	0.92	0.91	0.91	0.91
	Ours	**0.64**	**0.71**	**0.67**	**0.61**	**0.81**	**0.69**	**0.98**	**0.96**	**0.96**	**0.93**	**0.98**	**0.96**
LLaVA-v1.6-7B	APS	0.70	0.34	0.61	0.57	**0.82**	0.61	0.96	0.87	0.91	0.92	0.90	0.91
	LAC	0.68	0.48	0.56	0.50	0.69	0.58	0.92	0.81	0.87	0.87	0.87	0.87
	VLM-Bench	0.69	0.41	0.59	0.54	0.76	0.60	0.94	0.84	0.89	0.90	0.89	0.89
	Ours	**0.70**	**0.56**	**0.62**	**0.59**	0.76	**0.65**	**0.97**	**0.93**	**0.94**	**0.94**	**0.97**	**0.95**
LLMs		**HSwg**	**HDial**	**CQA**	**HSum**	**MMLU**	**Avg.**	**HSwg**	**HDial**	**CQA**	**HSum**	**MMLU**	**Avg.**
Yi-34B	APS	0.91	0.51	0.84	0.56	0.59	0.68	0.97	0.73	0.94	0.79	0.88	0.86
	LAC	0.95	0.57	0.93	0.42	0.68	0.71	0.97	0.71	0.93	0.75	0.86	0.85
	LLM-Bench	0.93	0.54	0.89	0.49	0.64	0.70	0.97	0.72	0.94	0.77	0.87	0.86
	Ours	**0.97**	**0.70**	**0.96**	**0.62**	**0.74**	**0.80**	**1.00**	**0.96**	**1.00**	**0.93**	**0.97**	**0.97**
Qwen-14B	APS	0.84	0.53	0.79	0.26	0.64	0.61	0.98	0.83	0.97	0.63	0.86	0.85
	LAC	0.92	0.48	0.91	0.13	0.54	0.60	0.97	0.80	0.97	0.57	0.82	0.83
	LLM-Bench	0.88	0.51	0.85	0.20	0.59	0.61	0.98	0.82	0.97	0.60	0.84	0.84
	Ours	**0.94**	**0.62**	**0.93**	**0.35**	**0.65**	**0.70**	**0.99**	**0.93**	**0.99**	**0.71**	**0.95**	**0.92**
Qwen-7B	LLM-Bench	0.51	0.35	0.70	0.22	0.47	0.45	0.68	0.75	0.89	0.50	0.73	0.71
	Ours	**0.64**	**0.50**	**0.80**	**0.44**	**0.61**	**0.60**	**0.93**	**0.92**	**0.98**	**0.68**	**0.95**	**0.89**

meets this requirement (set as 90% coverage) across all evaluated models and datasets, demonstrating robust statistical calibration. In contrast, APS achieves high empirical coverage but does so by significantly inflating prediction sets, reflecting overconservatism rather than meaningful uncertainty quantification. LAC, while aiming for compactness, often undercovers, violating the intended guarantee. This contrast highlights a core trade-off in conformal inference: ensuring valid coverage while maintaining informative and actionable prediction sets. CAP strikes this balance by adaptively selecting instance-wise risk levels through its learned abstention policy.

As demonstrated in Table 3, CAP achieves a favorable balance between predictive accuracy and set compactness. It substantially reduces prediction set sizes relative to APS—often by over 25% without sacrificing coverage, indicating that APS's large sets are unnecessarily conservative. At the same time, CAP consistently outperforms LAC in accuracy across nearly all datasets and models, while avoiding LAC's frequent undercoverage due to overly narrow sets. Notably, CAP maintains calibration closely aligned with the nominal coverage target, with no systematic bias toward over- or under-coverage, reflecting a robust generalization capability across both VLM and LLM settings. This highlights CAP's ability to learn context-sensitive abstention thresholds that retain statistical validity.

Table 2: Coverage (%) comparison of CAP (ours), Benchmarking LLMs via Uncertainty Quantification Ye et al. (2024), and Uncertainty-Aware Evaluation Kostumov et al. (2024) across VLMs and LLMs. CAP consistently meets the 90% coverage guarantee while maintaining tighter set sizes.

Model	Method	VLMs (Coverage %) ↑					LLMs (Coverage %) ↑				
		MMB	OOD	SQA	SB	Avg.	HSwg	HDial	CQA	MMLU	Avg.
LLaVA-34B/Qwen-7B	Bench	**94.49**	93.14	**93.37**	**93.02**	**93.57**	91.53	**92.43**	94.02	**92.79**	**92.69**
	Ours	93.97	**93.25**	93.07	91.41	93.43	**91.96**	91.70	**95.68**	91.32	92.16
LLaVA-13B/Qwen-14B	Bench	94.59	**93.60**	**94.28**	**93.60**	**94.02**	**95.29**	**91.92**	95.2	**92.58**	**93.75**
	Ours	**95.57**	92.48	92.06	90.67	93.18	94.88	90.96	**95.66**	91.62	92.68
LLaVA-7B/Yi-34B	Bench	**93.86**	**93.5**	**93.85**	**93.47**	**93.67**	**96.89**	**92.63**	**97.04**	**93.54**	**94.16**
	Ours	92.96	91.63	90.49	91.23	91.94	96.48	92.56	96.40	93.34	93.92

Table 3: Accuracy (%) and set size comparison of CAP (Ours) with Benchmarking LLMs via Uncertainty Quantification Ye et al. (2024), and Uncertainty-Aware Evaluation Kostumov et al. (2024). CAP achieves the highest average accuracy while maintaining balanced set sizes. Highest accuracies are in **bold**, and well-calibrated set sizes are underlined.

Models	Method	Accuracy (%) ↑						SS ↓					
VLMs		**MMB**	**OOD**	**SQA**	**SB**	**AI2D**	**Avg.**	**MMB**	**OOD**	**SQA**	**SB**	**AI2D**	**Avg.**
LLaVA-v1.6-34B	Bench	87.24	86.94	83.95	81.56	82.88	84.52	1.95	1.49	2.01	2.12	2.06	1.93
	Ours	**88.57**	**88.19**	**86.46**	**81.64**	**88.36**	**86.64**	1.65	1.62	1.84	1.99	2.18	1.86
LLaVA-v1.6-13B	Bench	76.75	72.93	70.56	70.37	73.67	72.85	2.34	2.18	2.45	2.49	2.33	2.36
	Ours	**82.66**	**80.79**	**79.08**	**77.50**	**84.87**	**81.38**	2.62	2.23	2.18	2.28	2.31	2.32
LLaVA-v1.6-7B	Bench	75.56	73.7	65.86	69.06	69.75	70.78	2.37	2.34	2.45	2.53	2.37	2.41
	Ours	**82.19**	**81.20**	**75.34**	**76.34**	**81.83**	**79.38**	1.99	2.30	2.19	2.35	2.37	2.24
LLMs		**HSwg**	**HDial**	**CQA**	**HSum**	**MMLU**	**Avg.**	**HSwg**	**HDial**	**CQA**	**HSum**	**MMLU**	**Avg.**
Yi-34B	Bench	94.56	83.58	95.07	81.09	80.54	87.37	2.01	1.72	1.79	1.62	2.21	1.88
	Ours	**96.17**	**85.56**	**96.12**	**83.09**	**82.90**	**88.77**	1.48	2.07	1.57	1.85	2.12	1.82
Qwen-14B	Bench	91.00	73.90	91.52	49.33	64.25	74.00	2.02	1.94	1.74	2.37	2.80	2.17
	Ours	**94.02**	**83.09**	**94.32**	**57.59**	**76.13**	**81.03**	1.38	1.87	1.33	2.38	2.55	1.90
Qwen-7B	Bench	63.70	64.04	83.89	32.53	55.21	59.87	2.28	2.51	2.15	2.92	3.26	2.63
	Ours	**73.79**	**75.81**	**90.06**	**47.75**	**72.25**	**71.93**	2.61	2.88	1.92	2.57	3.18	2.63

4.2.3. Accuracy and Informativeness

We evaluate accuracy alongside prediction set size, a key proxy for informativeness. As shown in Table 3, CAP consistently achieves the highest average accuracy across all models and datasets, outperforming both APS and LAC. While APS often matches or approaches CAP in accuracy, it does so by generating excessively large prediction sets—up to 3× the size of CAP's—thereby reducing informativeness. LAC, in contrast, produces smaller sets but often fails to maintain statistical validity, resulting in lower accuracy. CAP uniquely balances these objectives: it adaptively calibrates prediction sets that are significantly more compact than APS while preserving the coverage guarantees absent in LAC. This demonstrates CAP's ability to jointly optimize for correctness and confidence, capturing fine-grained uncertainty that static baselines fail to model in isolation.

Table 4: Evaluation of Expected Calibration Error (ECE): Comparative analysis of the proposed CAP framework (Ours) with Benchmarking LLMs via Uncertainty Quantification Ye et al. (2024), Uncertainty-Aware Evaluation Kostumov et al. (2024) as well as with standard LAC Sadinle et al. (2019), and APS Romano et al. (2020) methods. Evidently, CAP achieves significantly lower ECE values, in **bold**, compared to baselines.

Method	VLM Datasets (ECE ↓)						LLM Datasets (ECE ↓)					
	MMB	OOD	SQA	SB	AI2D	Avg.	HSwg	HDial	CQA	HSum	MMLU	Avg.
	LLaVA-v1.6-34B						**Yi-34B**					
APS	0.13	0.13	0.21	0.14	0.24	0.17	0.11	0.32	0.16	0.22	0.25	0.21
LAC	0.07	0.11	0.11	0.13	0.16	0.11	0.05	0.27	0.10	0.19	0.17	0.16
Bench	0.10	0.12	0.16	0.14	0.20	0.14	0.08	0.30	0.13	0.21	0.21	0.18
Ours	**0.01**	**0.03**	**0.03**	**0.03**	**0.04**	**0.03**	**0.05**	**0.15**	**0.10**	**0.05**	**0.03**	**0.08**
	LLaVA-v1.6-13B						**Qwen-14B**					
APS	0.16	0.22	0.19	0.16	0.27	0.20	0.09	0.20	0.10	0.29	0.27	0.19
LAC	0.13	0.17	0.16	0.18	0.19	0.17	0.02	0.16	0.03	0.34	0.23	0.15
Bench	0.06	0.02	0.1	0.11	0.03	0.06	0.05	0.18	0.06	0.31	0.25	0.17
Ours	**0.02**	**0.02**	**0.04**	**0.06**	**0.03**	**0.03**	**0.01**	**0.03**	**0.03**	**0.13**	**0.02**	**0.04**
	LLaVA-v1.6-7B						**Qwen-7B**					
APS	0.16	0.24	0.21	0.17	0.26	0.21	0.35	0.30	0.23	0.41	0.40	0.34
LAC	0.13	0.20	0.19	0.18	0.20	0.18	0.32	0.27	0.15	0.44	0.35	0.30
Bench	0.07	0.04	0.10	0.11	0.03	0.07	0.33	0.28	0.19	0.42	0.37	0.32
Ours	**0.04**	**0.03**	**0.05**	**0.06**	**0.01**	**0.04**	**0.08**	**0.03**	**0.08**	**0.14**	**0.05**	**0.07**

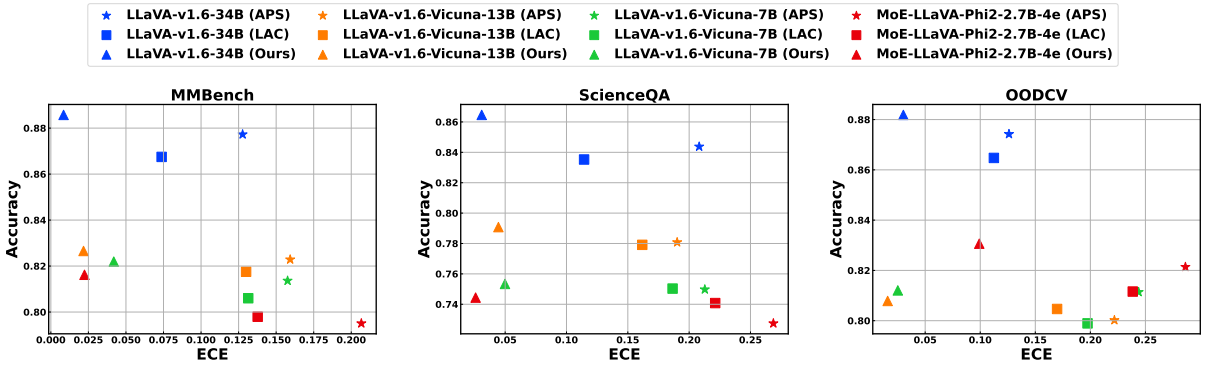

Figure 1: Accuracy vs. ECE for CAP against baseline across five VLM datasets. CAP demonstrates substantially lower ECE, placing results closer to the ideal upper-left region.

4.2.4. Calibration Performance

We assess calibration quality using ECE, which quantifies the discrepancy between predicted confidence and observed accuracy. As shown in Table 4, the proposed CAP framework substantially outperforms APS and LAC across both vision-language and language-only models. Averaged over all model-dataset pairs, CAP reduces ECE by **82.9%** relative to APS and **74.8%** relative to LAC, indicating significantly improved confidence alignment. Notably, CAP achieves this without compromising accuracy—and often improving it—demonstrating that its uncertainty estimates are not only conservative but also sharp (Table 3). Visualized

in Figure 1, CAP consistently lies in the upper-left quadrant of the accuracy-ECE frontier, indicating both high correctness and strong reliability.

Extended datasets and evaluation (supplementary). In addition to the ten main MCQA benchmarks, we include supplementary results in **??** on extended datasets and evaluation settings. These include detailed accuracy–calibration curves, ablations across model scales, and extended comparisons on selective generation and hallucination detection tasks. All supplementary datasets follow the same MCQA formatting and calibration–test protocol as here. This extended evaluation further demonstrate the robustness of CAP across diverse tasks and models, demonstrating its consistent ability to improve abstention quality, calibration, and factuality in both multiple-choice and free-form settings.

4.2.5. Extension to Free-Form Generation

While our primary experiments focus on MCQA tasks to ensure controlled and standardized evaluation, the CAP framework is also applicable to mitigating hallucinations in free-form generation. In this setting, the goal shifts from abstaining on a discrete choice to controlling the factuality and informativeness of a generated text. We adapt CAP by conceptualizing the problem through the lens of sub-claim management, an approach validated in recent work on conformal factuality Mohri and Hashimoto. Here, an initial free-form response from a base LLM is first decomposed into a set of atomic sub-claims. The role of the CAP policy, $\pi_\theta(\alpha|x)$, is to learn an optimal risk level α that translates to a retention threshold for these sub-claims. A higher risk level corresponds to a more aggressive removal of less confident claims. The utility function (Equation 3) is updated to reward factuality (number of correct claims retained) while penalizing loss of informativeness (number of total claims removed). This allows CAP to dynamically manage the trade-off between generating a detailed response and ensuring its factuality on a per-instance basis.

Original Generation (Yi-34B)	**CAP-Controlled**
Zamfir Ralli-Arbore (1848-1933) was a Romanian political activist and historian from Bessarabia, who spent much of his life in exile. As a member of the National Liberal Party, he campaigned for the union of his native region with the Kingdom of Romania, and was a prominent opponent of Russian and Soviet policies. He was also a noted historian, specializing in the history of the Moldavia and Wallachia during the Middle Ages.	Zamfir Ralli-Arbore, born in 1848, was a Romanian political activist from Bessarabia. He was a noted historian, specializing in the history of the Moldavia and Wallachia. He passed away in 1933.

Figure 2: An example of CAP's application to free-form generation on the FActScore dataset. The original text contains hallucinated details (highlighted in red) which are successfully omitted by the CAP policy, preserving only the verified factual information.

Table 5 presents a comparative analysis using the experimental setup from Mohri and Hashimoto. We compare against their Conformal Factuality (CF) framework on the FActScore, Natural Questions (NQ), and MATH datasets. The metric shown is the percentage of sub-claims retained while achieving a target factuality of 90%. As a baseline, the unmodified base model (Yi-34B) has a much lower initial factuality.

The results indicate that CAP consistently outperforms the CF baseline. For instance, on the challenging FActScore dataset, CAP retains up to 12.7% more information. This superiority stems from CAP's learned, instance-specific policy, which is more efficient at identifying and removing only the truly non-factual claims, whereas the static threshold

Table 5: Factuality of the base Yi-34B model and downstream claim-retention rate after post-processing. CAP preserves markedly more content than Conformal Factuality (CF) while satisfying the 90 % factuality target. "Δ" denotes the absolute retention gain of CAP.

Dataset	Base Fact. (%)	CF Ret. (%)	CAP Ret. (%)	Δ
FActScore	34.1 (100)	35.5	**48.2**	+12.7
NQ	76.5 (100)	74.8	**83.1**	+8.3
MATH	74.2 (100)	89.5	**94.3**	+4.8

in standard conformal methods can be overly conservative, removing correct-but-uncertain claims. A qualitative example of this is shown in Figure 2. Ultimately, this extension demonstrates that CAP's core mechanism—learning a utility-maximizing abstention policy—is a powerful and flexible approach for risk management that generalizes effectively from structured classification to open-ended generative tasks.

5. Conclusion and Future Work

We introduced a RL-based framework to adaptively configure conformal prediction thresholds for selective abstention in LLMs and VLMs. By learning to adjust the decision boundary between point predictions, set-valued outputs, and abstention, our method overcomes the rigidity of static conformal approaches, enabling instance-specific coverage–informativeness trade-offs and improved calibration. Extensive evaluations across diverse tasks demonstrate that CAP consistently outperforms APS and LAC. CAP improves hallucination detection (AUROC) by up to 22.2% and selective generation (AUARC) by 21.2%, reduces calibration error by over 70%, and produces tighter prediction sets while preserving valid coverage guarantees. It achieves higher accuracy than LAC and more compact, informative prediction sets than APS, establishing a new standard for utility-aware abstention in foundation models.

However, integrating CP with RL introduces challenges. Learned policies may overfit to calibration data, leading to biased abstention behaviors or compromising CP's theoretical guarantees. CAP also introduces additional parameters and depends on well-tuned reward functions that may require adaptation to new data distributions. To mitigate these risks, future work can explore distribution-aware regularization, policy validation using out-of-sample data, and constrained reward design.

References

Sara Abdali, Richard Anarfi, CJ Barberan, and Jia He. Securing large language models: Threats, vulnerabilities and responsible practices. *arXiv preprint arXiv:2403.12503*, 2024.

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

John Cherian, Isaac Gibbs, and Emmanuel Candes. Large language model validity via enhanced conformal prediction methods. *Advances in Neural Information Processing Systems*, 37:114812–114842, 2024.

C Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on information theory*, 16(1):41–46, 1970.

Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, 2006.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. *International Conference on Machine Learning*, pages 1050–1059, 2016.

Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, et al. A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, 56(Suppl 1):1513–1589, 2023.

Yonatan Geifman and Ran El-Yaniv. Selectivenet: A deep neural network with an integrated reject option. In *International conference on machine learning*. PMLR, 2019.

Kilian Hendrickx, Lorenzo Perini, Dries Van der Plas, Wannes Meert, and Jesse Davis. Machine learning with a reject option: A survey. *Machine Learning*, 113(5), 2024.

Yuheng Huang, Jiayang Song, Zhijie Wang, Shengming Zhao, Huaming Chen, Felix Juefei-Xu, and Lei Ma. Look before you leap: An exploratory study of uncertainty measurement for large language models. *arXiv preprint arXiv:2307.10236*, 2023.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Vasily Kostumov, Bulat Nutfullin, Oleg Pilipenko, and Eugene Ilyushin. Uncertainty-aware evaluation for vision-language models. *arXiv preprint arXiv:2402.14418*, 2024.

Ranganath Krishnan, Piyush Khanna, and Omesh Tickoo. Enhancing trust in large language models with uncertainty-aware fine-tuning. *arXiv preprint arXiv:2412.02904*, 2024.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in Neural Information Processing Systems*, 30, 2017.

Minjae Lee, Kyungmin Kim, Taesoo Kim, and Sangdon Park. Selective generation for controllable language models. *Advances in Neural Information Processing Systems*, 37: 50494–50527, 2024.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

Christopher Mohri and Tatsunori Hashimoto. Language models with conformal factuality guarantees, 2024. *URL https://arxiv. org/abs/2402.10978*.

Jishnu Mukhoti, Andreas Kirsch, Joost van Amersfoort, Philip HS Torr, and Yarin Gal. Deep deterministic uncertainty: A new simple baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24384–24394, 2023.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29, 2015.

Andrea Pugnana, Lorenzo Perini, Jesse Davis, and Salvatore Ruggieri. Deep neural network benchmarks for selective classification. *arXiv preprint arXiv:2401.12708*, 2024.

Victor Quach, Adam Fisch, Tal Schuster, Adam Yala, Jae Ho Sohn, Tommi S Jaakkola, and Regina Barzilay. Conformal language modeling. *arXiv preprint arXiv:2306.10193*, 2023.

Yaniv Romano, Matteo Sesia, and Emmanuel Candes. Classification with valid and adaptive coverage. *Advances in Neural Information Processing Systems*, 33:3581–3591, 2020.

Mauricio Sadinle, Jing Lei, and Larry Wasserman. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525): 223–234, 2019.

Renukanandan Tumu, Matthew Cleaveland, Rahul Mangharam, George Pappas, and Lars Lindemann. Multi-modal conformal prediction regions by optimizing convex shape templates. In *6th Annual Learning for Dynamics & Control Conference*. PMLR, 2024.

Neeraj Varshney, Swaroop Mishra, and Chitta Baral. Investigating selective prediction approaches across several tasks in iid, ood, and adversarial settings. *arXiv preprint arXiv:2203.00211*, 2022.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*, volume 29. Springer, 2005.

Yasin Abbasi Yadkori, Ilja Kuzborskij, David Stutz, András György, Adam Fisch, Arnaud Doucet, Iuliya Beloshapka, Wei-Hung Weng, Yao-Yuan Yang, Csaba Szepesvári, et al. Mitigating llm hallucinations via conformal abstention. *arXiv preprint arXiv:2405.01563*, 2024.

Fanghua Ye, Mingming Yang, Jianhui Pang, Longyue Wang, Derek Wong, Emine Yilmaz, Shuming Shi, and Zhaopeng Tu. Benchmarking llms via uncertainty quantification. *Advances in Neural Information Processing Systems*, 37:15356–15385, 2024.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *International conference on machine learning*, pages 12697–12706. PMLR, 2021.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36: 46595–46623, 2023.