# Geometry-Calibrated Conformal Abstention for Language Models

**Rui Xu**[1] **Yi Chen**[2] **Sihong Xie**[1] **Hui Xiong**[1]

[1] The Hong Kong University of Science and Technology (Guangzhou) [2] Jilin University

rxu233@connect.hkust-gz.edu.cn chenyi1522@mails.jlu.edu.cn
sihongxie@hkust-gz.edu.cn xionghui@hkust-gz.edu.cn

## Abstract

When language models lack relevant knowledge for a given query, they frequently generate plausible responses that can be hallucinations, rather than admitting being agnostic about the answer. Retraining models to reward admitting ignorance can lead to overly conservative behaviors and poor generalization due to scarce evaluation benchmarks. We propose a post-hoc framework, *Conformal Abstention* (CA), adapted from conformal prediction (CP) to determine whether to abstain from answering a query. CA provides a finite-sample participation guarantee and a data-dependent lower bound on the probability that a generated response is correct given participation. The abstention decision relies on prediction confidence rather than the nonconformity scores used in CP, which are intractable for open-ended generation. To better align prediction confidence with the model's ignorance, we introduce a calibration strategy using representation geometry within the model to measure knowledge involvement in shaping the response. Experiments demonstrate that our method improves selective answering with 75% conditional correctness.

## 1 Introduction

Advances in model architectures [1] have endowed language models with broad linguistic competence. Leveraging the ability, large language models (LLMs) have become central components of artificial intelligence, serving as general-purpose engines for reasoning [2], knowledge retrieval [3], and decision support [4], with growing deployment in science [5], healthcare [6], and education [7]. Despite their remarkable capabilities, LLMs should refuse to respond in certain situations, such as queries involving safety [8] or privacy [9]. Beyond externally imposed constraints, LLMs must also recognize the limits of their own knowledge [10]. When faced with questions for which they do not know the answer, models tend to generate responses without reliable grounding, producing responses that appear plausible but are in fact hallucinated. This behavior raises persistent concerns about reliability in real-world deployments [11]. Importantly, this phenomenon is not solely due to limitations in model capacity, but is closely tied to commonly used training loss functions [12]. Most evaluation protocols reward models only for producing an answer, typically measuring performance through accuracy. Consequently, abstention (e.g., by responding "I don't know") receives zero or even negative rewards, while guessing always offers a nonzero chance of success, making guessing more rewarding than abstaining for all queries. Recent work suggests that rewarding appropriate expressions of ignorance, rather than penalizing them, may help mitigate this issue [13].

However, the scarcity of benchmarks that explicitly reward abstention causes models to overfit, which in turn degrades their generalization ability [14]. Moreover, excessively rewarding model refusal can induce indiscriminate abstention [12, 15]. As a result, promoting abstention through training-based interventions can be fragile. These considerations motivate a central question:

*Can LLMs abstain using a post-hoc mechanism?*

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

We propose **Conformal Abstention (CA)**, an approach grounded in conformal prediction (CP) [16] that enables LLMs to admit a lack of knowledge. Specifically, CA establishes two finite-sample theoretical results under the assumption of exchangeability: (i) a *participation guarantee*, bounding the probability that the model participates (i.e., an LLM provides an informative answer) rather than saying "I don't know"; and (ii) a *conditional-correctness lower bound* on answer accuracy given model participation. The abstention decision is based on prediction confidence, avoiding the nonconformity scores used in CP, which are empirically intractable for open-ended generation [17].

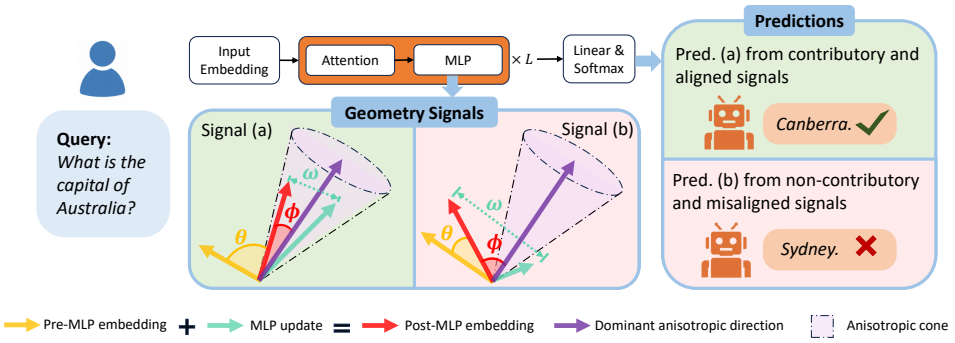



Figure 1: Token representation geometry captures how the model's internal knowledge shapes its response.

To derive prediction confidence that faithfully reflects when the model is truly uncertain, we analyze how internal knowledge influences the generation process through the trajectory of token representations across layers. Specifically, we focus on the update induced by the MLP, which is known to encode the model's factual knowledge. As illustrated in Figure 1, we quantify this influence using three geometric signals derived from the language model's hidden states.

1. **proximity** $\omega$, measuring the MLP's contribution via its distance to the post-MLP embedding;
2. **embedding rotation** $\theta$, capturing the extent to which the update alters the prediction direction;
3. **alignment** $\phi$, measuring consistency with the dominant anisotropic direction.

Intuitively, when the MLP update contributes strongly (small distance to the post-MLP embedding), induces a significant rotation, and aligns well with the dominant direction, the model is more likely to leverage useful internal knowledge, leading to a correct prediction (Figure 1 signal and pred. (a)). Conversely, weak contribution or misalignment indicates uncertainty, where the response is less reliable (Figure 1 signal and pred. (b)). We leverage these geometric signals to calibrate confidence.

Across six datasets, the proposed method consistently outperforms a broad set of strong baselines, achieving the highest average conditional correctness of 75.0% under conformal abstention.

## 2 Background

### 2.1 Conformal Prediction for Language Models

We begin with a brief review of conformal prediction (CP) [18]. Throughout, we use uppercase letters (e.g.,$X$) to denote random variables, lowercase letters (e.g., $x$) to denote their realizations, and script letters (e.g., $\mathcal{X}$) to denote sets, unless stated otherwise. Given a model $f : \mathcal{X} \to \mathcal{Y}$, a score function $g : \mathcal{X} \times \mathcal{Y} \to \mathcal{S} \subseteq \mathbb{R}$ outputs nonconformity scores to assess how data conforms to the model $f$, with lower values indicating better conformity. For classification tasks, it can be defined as the negative predicted probability of the true label. With calibration instances $\{(X_i, Y_i)\}_{i=1}^n$, split conformal prediction computes calibration nonconformity scores $S_i := g(X_i, Y_i)$ for $i = 1, \ldots, n$ [19]. Let $\tau$ be the $\lceil (1-\alpha)(n+1) \rceil / n$ quantile of $\{S_i\}_{i=1}^n$.[1] For a test instance $(X_{n+1}, Y_{n+1})$, the corresponding prediction set is $\mathcal{C}(X_{n+1}) := \{y \in \mathcal{Y} : g(X_{n+1}, y) \leq \tau\}$. Assuming exchangeability, we can ensure

$$\mathbb{P}\big(Y_{n+1} \in \mathcal{C}(X_{n+1})\big) = \mathbb{P}(S_{n+1} \leq \tau) \geq 1 - \alpha. \tag{1}$$

However, in autoregressive language generation, ground-truth answers can take semantically equivalent forms, making it intractable to define the score function $g$ via likelihood over all valid outputs.

[1] $\tau$ is the $\lceil (1-\alpha)(n+1) \rceil$-th smallest value among $\{S_i\}_{i=1}^n$.

To address this, prior work restricts generation to structured settings such as multiple-choice classification [20, 21, 22] or regression-based grading [23], which limits applicability to open-ended tasks. Another line of work approximates nonconformity via repeated sampling, but relies on the unrealistic assumption that the ground-truth answer is sampled [24, 25]. When this assumption fails, valid coverage can only be achieved if $1-\alpha$ is below the fraction of calibration instances where the correct answer appears [26, 27]. Alternatively, some approaches redefine correctness using semantic entailment [28] or admissibility [29], weakening the connection between conformal guarantees and true correctness, as such criteria may still admit incomplete or misleading responses.

Abstention has recently emerged as a promising alternative direction. For example, Yadkori et al. [30] adopt conformal risk control [31] to bound the joint probability of model participation (i.e., not abstaining) and incorrect prediction. However, since model participation is directly observable, the conditional probability of correctness given participation is more useful and practically relevant than the joint probability. Our theoretical framework derives a lower bound on this conditional probability.

### 2.2 Estimating Reliability from Internal Model Signals

Reliability estimation based on surface-level signals, such as prediction probabilities [32] or sampling-based consistency [33, 34], is often insufficient for LLMs. Recent work instead leverages internal signals, including attention spectra [35], hidden-state statistics [36], and embedding distances [37]. However, these approaches overlook how internal knowledge shapes the evolution of representations. When relevant knowledge is present, it should induce greater changes in hidden states. This suggests that the geometry of representation transformations provides a meaningful signal for reliability. Motivated by this hypothesis, we derive geometry-based signals from representation transformations and examine their association with prediction correctness. We then use these signals to calibrate the model's confidence.

## 3 Conformal Abstention

### 3.1 Participation Guarantee

Consider a large language model $f$ and input query $x$. Let $\widehat{y} := f(x) = (\widehat{y}[1], ..., \widehat{y}[N])$ denote the generated output consisting of $N$ tokens. We associate each query $x$ with an *uncertainty score* $u := r(\widehat{y}|x)$ to reflect the reliability of the model's response $\widehat{y}$. Here, $r$ denotes a user-specified uncertainty scoring function that outputs a scalar correlated with prediction correctness; several such functions have been proposed in prior work [32, 38]. A natural choice is to derive uncertainty from the model's confidence in its own generation. For instance, denoting $p(\cdot|x)$ the conditional token distribution induced by the model $f$ given $x$, we employ perplexity and define

$$r^{\text{perp}}(\widehat{y}|x) = \exp\left(-\frac{1}{N}\sum\nolimits_{t=1}^{N} \log p(\widehat{y}[t] \mid x, \widehat{y}[<t])\right) \tag{2}$$

where $\widehat{y}[<t]$ is the prefix up to the token at $t-1$.

We then introduce a threshold on the uncertainty score and abstain from prediction on inputs whose scores exceed this threshold. Formally, given $n$ calibration samples $\{(X_i, Y_i)\}_{i=1}^n$, we obtain response $\widehat{Y}_i := f(X_i)$ and compute uncertainty score $U_i := r(\widehat{Y}_i|X_i)$, for $i = 1, \ldots, n$. Let $\tau$ be the $\lceil (1-\alpha)(n+1) \rceil / n$ quantile of $\{U_i\}_{i=1}^n$:

$$\tau := \text{Quantile}\left(\lceil (1-\alpha)(n+1) \rceil / n, \{U_i\}_{i=1}^n\right). \tag{3}$$

A *kept region* is defined as $\mathcal{K} := \{x : r(x) \leq \tau\}$. The model participates (e.g., generates an informative answer) instead of saying "I do not know" for samples in $\mathcal{K}$. If a test sample $(X_{n+1}, Y_{n+1})$ is jointly exchangeable with the $n$ calibration samples, and $U_1, ..., U_n$ are almost surely distinct, standard conformal arguments [16] yield a *participation guarantee*:

$$\mathbb{P}(X_{n+1} \in \mathcal{K}) = \mathbb{P}(U_{n+1} \leq \tau) \in [1-\alpha, 1-\alpha+1/(n+1)). \tag{4}$$

### 3.2 Conditional Correctness Lower Bound

To determine whether a prediction $\widehat{y}$ for a prompt $x$ is correct, we define an *evaluation function* $\xi$ with respect to the ground-truth answer $y$, whose realization is detailed in Appendix B. Specifically,

we have

$$\xi(x, y, \widehat{y}) := \begin{cases} 1, & \text{if } \widehat{y} \text{ is correct for } x \text{ with respect to } y, \\ 0, & \text{otherwise.} \end{cases}$$

For notational clarity, let $J := \xi(X, Y, \widehat{Y})$ be the random variable indicating prediction correctness. We aim to bound the probability that a prediction is correct conditioned on model participation, i.e.,

$$\mathbb{P}(J_{n+1} = 1 \mid X_{n+1} \in \mathcal{K}).$$

To derive a lower bound on conditional correctness, we establish that the test uncertainty score and correctness indicator $(U_{n+1}, J_{n+1})$ are exchangeable with those of the calibration samples. This follows from Theorem 3.1 showing that exchangeability is preserved under symmetric deterministic transformations, leading directly to Lemma 3.2.

**Theorem 3.1** (Exchangeability Preservation). *[39] Let $(Z_1, ..., Z_{n+1}) \in \mathcal{Z}^{n+1}$ be a vector of exchangeable random variables. Fix a transformation $g : \mathcal{Z}^{n+1} \to (\mathcal{Z}')^{n+1}$. If for each permutation $\pi_1$, there exists a permutation $\pi_2$ such that $\pi_1 g(z_1, ..., z_{n+1}) = g(\pi_2(z_1, ..., z_{n+1}))\ \forall (z_1, ..., z_{n+1}) \in \mathcal{Z}^{n+1}$, then $(Z'_1, ..., Z'_{n+1}) := g(Z_1, ...Z_{n+1})$ is a vector of exchangeable random variables.*

**Lemma 3.2** (Exchangeability of Uncertainty and Correctness Pairs). *Let $(X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1})$ be exchangeable random variables representing questions and ground-truth answers. Define*

$$U_i := r(f(X_i)|X_i), \quad J_i := \xi(X_i, Y_i, f(X_i)), \quad \text{for each} \quad i = 1, \ldots, n+1,$$

*where $r(\cdot)$ is an uncertainty scoring function, $\xi(\cdot)$ is a correctness evaluation function, and $f(\cdot)$ is a deterministic language model. Then $(U_1, J_1), \ldots, (U_{n+1}, J_{n+1})$ are exchangeable.*

Lemma 3.2 (proved in Appendix A.1) allows reasoning about the subset of uncertainty scores corresponding to correctly predicted samples. Specifically, conditional on the event that the test prediction is correct ($J_{n+1} = 1$), the test uncertainty score $U_{n+1}$ is exchangeable with the set of calibration uncertainty scores for correctly predicted samples $\mathcal{U} := \{ U_i : J_i = 1,\ i = 1, \ldots, n \}$.

Let $c$ denote the cardinality of $\mathcal{U}$, i.e., $c := |\mathcal{U}|$. The maximal coverage level associated with $\mathcal{U}$ that are no greater than the threshold $\tau$ is given by

$$1 - \beta \ := \ \sup\{q \in (0, 1) : \text{Quantile}(\lceil q(c+1) \rceil / c, \mathcal{U}) \leq \tau\}.$$

Let $\sigma := \text{Quantile}\big(\lceil (1-\beta)(c+1) \rceil / c, \mathcal{U}\big)$ be the quantile value of $\mathcal{U}$ corresponding to the maximal coverage level $1 - \beta$ below $\tau$. By construction, we have $\sigma \leq \tau$, and hence

$$\mathbb{P}(X_{n+1} \in \mathcal{K} \mid J_{n+1} = 1) = \mathbb{P}(U_{n+1} \leq \tau \mid J_{n+1} = 1) \geq \mathbb{P}(U_{n+1} \leq \sigma \mid J_{n+1} = 1) \geq 1 - \beta, \quad (5)$$

providing a lower bound on the probability that a correct prediction is retained (i.e., not abstained).

By exchangeability, the test prediction is correct with probability $\Pr(J_{n+1} = 1) = \mathbb{E}[c/n]$, where the observed calibration accuracy $c/n$ is an estimate. Finally, combining the lower bound in Eq. (5) with the upper bound in Eq. (4) via Bayes' rule, the *conditional correctness lower bound* follows

$$\mathbb{P}(J_{n+1}{=}1 \mid X_{n+1} \in \mathcal{K}){=}\frac{\mathbb{P}(X_{n+1} \in \mathcal{K} \mid J_{n+1}{=}1)\mathbb{P}(J_{n+1}{=}1)}{\mathbb{P}(X_{n+1} \in \mathcal{K})} \geq \frac{(1-\beta) \cdot \mathbb{E}[c/n]}{1 - \alpha + 1/(n+1)}, \quad (6)$$

which bounds the probability that a prediction is correct, conditioned on model participation.

We verify the participation guarantee and the conditional-correctness lower bound on Simple-Questions-Wikidata [40] using Gemma-3-4B-Instruct [41], with perplexity in Eq. (2) as the uncertainty score. Figure 2 shows that the empirical participation rates satisfy their theoretical guarantees, while empirical conditional correctness exceeds the corresponding lower-bound estimates across a range of thresholds $\tau$, yielding positive gaps.

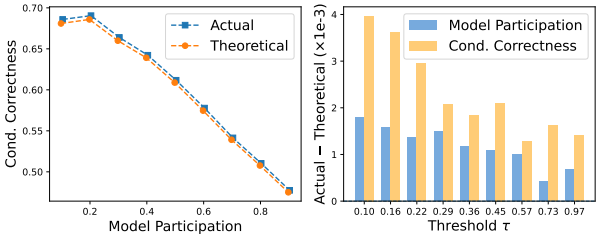



Figure 2: Participation and conditional correctness are ensured.

To achieve high conditional correctness, it is crucial to increase $1 - \beta$ in Eq. (6), since $1 - \alpha$ is user-specified, whereas $\mathbb{E}[c/n]$ corresponds to the model's population accuracy. Eq. (5) shows that $1 - \beta$ lower bounds the retention rate of correct predictions. This quantity increases when correct predictions are assigned lower uncertainty scores. Equivalently, improving conditional correctness requires assigning higher confidence to correct predictions, thereby increasing their probability of being retained.

## 4 Calibration via Representation Geometry

We posit that a model's prediction is more likely to be correct when its internal knowledge actively shapes the representations from which the prediction tokens are decoded. Since MLPs serve as the primary locus of memorized knowledge [42], we analyze whether their updates induce substantive geometric transformations of the representations across layers, as opposed to simple passive propagation. These geometry-based signals are subsequently used to calibrate prediction confidence.

### 4.1 Preliminary

Consider a language model composed of $L$ stacked transformer layers. Let a query $x = (x[1], \ldots, x[T])$ of $T$ tokens have hidden states $e^l = (e^l[1], \ldots, e^l[T]) \in \mathbb{R}^{d \times T}$ produced by layer $l$, with $e^0$ as the input embeddings. Each layer updates the representations as

$$\tilde{e}^l = e^{l-1} + \mathrm{Attn}^l(e^{l-1}), \quad e^l = \tilde{e}^l + \mathrm{MLP}^l(\tilde{e}^l), \tag{7}$$

where $\mathrm{Attn}^l(\cdot)$ and $\mathrm{MLP}^l(\cdot)$ denote the self-attention and MLP transformations at layer $l$, respectively, and $\tilde{e}^l$ is the intermediate representation after self-attention. However, the changes in representations do not disentangle whether they result from contextual understanding of the query or from knowledge injection, necessitating an explicit assessment of their origin.

### 4.2 Knowledge Contribution

Layer-wise knowledge contribution to representation changes is derived by aggregating information flow [43] between self-attention and MLP blocks.

**Information flow within language models** Assume that a language model uses multi-head self-attention with $H$ heads, each of dimension $D_H = D/H$. For head $h \in \{1, ..., H\}$ at layer $l$, the self-attention block produces an attention matrix $A^{lh} \in \mathbb{R}^{T \times T}$, and the corresponding value matrix is denoted $W_V^{lh} \in \mathbb{R}^{D_H \times D}$. Each layer has a projection matrix $W_O^l \in \mathbb{R}^{D \times D}$, which can be partitioned into $H$ submatrices $W_O^{lh} \in \mathbb{R}^{D \times D_H}$. With these notations, we can quantify the component from the source token embedding $e^{l-1}[s]$ to the self-attention output of the target token $\tilde{e}^l[t]$ by

$$a(t,s) := \mathbf{1}_{\{s=t\}} e^{l-1}[s] + \sum\nolimits_{h=1}^{H} W_O^{lh} A^{lh}[t,s] W_V^{lh} e^{l-1}[s],$$

where $\mathbf{1}_{\{s=t\}}$ is an indicator function that equals 1 if $s = t$ and 0 otherwise. Thereby, the target token embedding after self-attention transformation can be written as $\tilde{e}^l[t] = \sum_{s=1}^{T} a(t,s)$.

**Definition 4.1** (Normalized Proximity)**.** [43] Let $z \in \mathbb{R}^D$ admit a decomposition $z = \sum_{i=1}^{n} z_i$. The normalized proximity of component $z_i$ and $z$ follows

$$\mathrm{prox}\,(z, z_i) := \frac{\max(0, ||z||_1 - ||z - z_i||_1)}{\sum_{j=1}^{n} \max(0, ||z||_1 - ||z - z_j||_1)}. \tag{8}$$

High proximity indicates a strong contribution of $z_i$ to $z$. Based on Definition 4.1, we denote the attention contribution at layer $l$ by $C^l_{\mathrm{Attn}}$, with elements

$$C^l_{\mathrm{Attn}}[t,s] := \mathrm{prox}(\tilde{e}^l[t], a(t,s)). \tag{9}$$

Due to autoregressive attention, each token can only depend on itself and its predecessors. Hence, attention matrix $A^{lh}$ satisfies $A^{lh}[t,s] = 0$ if $s > t$. As a result, $C^l_{\mathrm{Attn}}$ is lower-triangular with entries above the diagonal equal to zero.

Since MLP blocks act independently on each token, the hidden state update by MLP can be written as $e^l[t] = \tilde{e}^l[t] + m^l[t]$, where $m^l[t] := \mathrm{MLP}^l(\tilde{e}^l[t])$. Using Definition 4.1, we define the MLP contribution matrix at layer $l$ as $C^l_{\mathrm{MLP}}$. Its diagonal entries are given by

$$C^l_{\mathrm{MLP}}[t,t] := \mathrm{prox}(e^l[t], m^l[t]), \tag{10}$$

while all off-diagonal entries are zero: $C^l_{\mathrm{MLP}}[t,s] := 0, \quad \forall s \neq t$, reflecting the absence of cross-token interactions on MLP operations. The information flow and contribution matrices for layer $l$ are visualized in Figure 3(a).

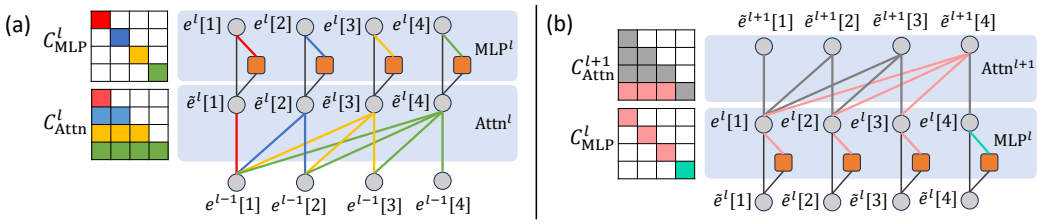



Figure 3: (a) Information flow and contribution matrices of layer $l$. Edge colors in the flow graph correspond to the different values in the contribution matrices (no edges means zero contribution). (b) Knowledge contribution from the MLP block at layer $l$. The **cyan edge and entry** indicate direct contribution to the last input token, corresponding to Eq. (11), while the **pink edges and entries** denote propagated contributions through subsequent self-attention, as described in Eq. (12).

**Knowledge Contribution** We use the matrices $C^l_{\text{Attn}}$ and $C^l_{\text{MLP}}$ for $l = 1, ..., L$ to measure how the knowledge stored in MLP blocks shapes the prediction probability of the generated token at position $T+1$ throughout layers. Figure 3 (b) shows that the gap between $\tilde{e}^l[T]$ and $e^l[T]$ is induced by the MLP update to the last query token, $m^l[T]$, so we define the direct knowledge contribution as

$$\omega^l_{\text{dir}} := C^l_{\text{MLP}}[T, T]. \tag{11}$$

The effect of layer $l$'s MLP on earlier tokens is not reflected in $e^l[T]$ but propagates via self-attention to $\tilde{e}^{l+1}[T]$ in the next layer. Thus, the propagated knowledge contribution is

$$\omega^l_{\text{prop}} := \text{diag}\big(C^l_{\text{MLP}}\big)^\top_{1:T-1} C^{l+1}_{\text{Attn}}[T, 1:T-1], \tag{12}$$

where $\text{diag}(\cdot)_{1:T-1}$ denotes the vector formed by the first $T-1$ diagonal entries of its matrix argument.

Together, Eq. (11) and Eq. (12) capture how MLP-stored knowledge actively shapes the representation of the last query token. We denote the full contribution trajectory by

$$\Omega := (\omega^1_{\text{dir}}, \omega^1_{\text{prop}}, ..., \omega^{L-1}_{\text{dir}}, \omega^{L-1}_{\text{prop}}, \omega^L_{\text{dir}}) \in \mathbb{R}^{2L-1}. \tag{13}$$

In the final layer, MLP updates to earlier tokens do not influence the last query token through self-attention, so the propagated contribution vanishes, and $\omega^L_{\text{prop}}$ is not defined.

### 4.3 Embedding Rotation

A high knowledge contribution does not always change the model's decision, which occurs when the token with the highest predicted probability at one layer is replaced by another token in the next layer. Therefore, we aim to link representation geometry to these decision dynamics.

**Lemma 4.2** (Scale Invariance of Token Ranking). *Let $\varepsilon \in \mathbb{R}^D$ be a token embedding, and let $o := W_E \cdot \varepsilon \in \mathbb{R}^V$ be the corresponding logits, where $W_E$ is the unembedding matrix. Let $\pi$ denote the ranking such that $o[\pi(1)] \geq o[\pi(2)] \geq \cdots \geq o[\pi(V)]$. For any $\lambda > 0$, define $\varepsilon' = \lambda\varepsilon$ and $o' = W_E \cdot \varepsilon'$. Then the ranking is invariant under positive scaling:*

$$o'[\pi(1)] \geq o'[\pi(2)] \geq \cdots \geq o'[\pi(V)] \tag{14}$$

**Lemma 4.3** (Rotation Sensitivity of Token Ranking). *Let $\varepsilon, \varepsilon' \in \mathbb{R}^D$ be two unit token embeddings with logits $o = W_E \cdot \varepsilon$, $o' = W_E \cdot \varepsilon' \in \mathbb{R}^V$ under the unembedding matrix $W_E$. Denote by $v_* = \arg\max_v o[v]$ the predicted token under $o$, and let $\theta$ be the angle between $\varepsilon$ and $\varepsilon'$. If*

$$2\|W_E\|_2 \sin(\theta/2) \leq \min_{v \neq v_*}(o[v_*] - o[v]), \tag{15}$$

*the top token is preserved, i.e., $\arg\max_v o'[v] = v_*$.*

By Lemma 4.2 (proved in Appendix A.2), rescaling an embedding does not affect the model's prediction. In contrast, Lemma 4.3 (proved in Appendix A.3) shows that large rotations can violate Eq. (15) and change the model's decision. Motivated by these results, we track how the embedding of the last query token rotates across layers. Specifically, for layer $l$, let $\theta^l_{\text{dir}}$ be the angle between $\tilde{e}^l[T]$ and $e^l[T]$, and let $\theta^l_{\text{prop}}$ be the angle between $e^l[T]$ and $\tilde{e}^{l+1}[T]$. We define the rotation trajectory by

$$\Theta := (\theta^1_{\text{dir}}, \theta^1_{\text{prop}}, ..., \theta^{L-1}_{\text{dir}}, \theta^{L-1}_{\text{prop}}, \theta^L_{\text{dir}}) \in \mathbb{R}^{2L-1}. \tag{16}$$

The relation between direct knowledge contribution $\omega^l_{\text{dir}}$ and its corresponding embedding rotation $\theta^l_{\text{dir}}$ is shown in Figure 4, which can be similarly extended to propagated contributions. Overall, knowledge contribution and embedding rotation are not necessarily tightly coupled, as similar levels of contribution can lead to different degrees of rotational change depending on the local geometry.

**Hypothesis 1.** Correctness is associated with cases where strong knowledge contribution induces significant embedding rotation, reflecting an effective redirection of the representation with increased prediction confidence.

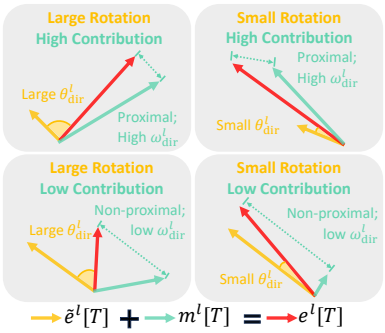



Figure 4: The relation between direct knowledge contribution and embedding rotation.

### 4.4 Anisotropy

Even with high knowledge contribution and large embedding rotation, the decision can still be incorrect due to misleading knowledge. Therefore, it is crucial to evaluate the likelihood that the resulting prediction is truly correct.

Transformer-based language models exhibit embeddings of training or in-distribution samples that concentrate within a narrow cone in the hidden space $\mathbb{R}^D$, inducing strong anisotropy [44]. This concentration yields well-defined dominant directions at each layer, which serve as canonical directions for typical in-distribution representations.

To distinguish between in-distribution and out-of-distribution behaviors, we define separate embedding distributions. Specifically, for layer $l$, let $\tilde{\eta}^{l,\text{in}}$ and $\eta^{l,\text{in}}$ denote the expectation of token embeddings produced by self-attention and MLP blocks for in-distribution inputs, and $\tilde{\eta}^{l,\text{out}}$ and $\eta^{l,\text{out}}$ denote their out-of-distribution counterparts.

**Hypothesis 2.** In-distribution queries induce hidden states that closely align with $\tilde{\eta}^{l,\text{in}}$ and $\eta^{l,\text{in}}$, remaining within the anisotropic cone and leading to more reliable predictions. In contrast, out-of-distribution queries deviate from this cone and shift toward $\tilde{\eta}^{l,\text{out}}$ and $\eta^{l,\text{out}}$. Such deviations indicate extrapolation beyond the model's training-induced representation region, resulting in uncertainty [45].

Analogous to Eq. (16), for alignment to dominant in-distribution directions, we define $\phi^{l,\text{in}}_{\text{dir}}$ as the angle between $\eta^l\text{in}$ and $e^l[T]$, and $\phi^{l,\text{in}}_{\text{prop}}$ as the angle between $\tilde{\eta}^{l+1}_{\text{in}}$ and $\tilde{e}^{l+1}[T]$. Similarly, we define $\phi^{l,\text{out}}_{\text{dir}}$ and $\phi^{l,\text{out}}_{\text{prop}}$ with respect to $\eta^{l,\text{out}}$ and $\tilde{\eta}^{l+1}_{\text{out}}$. The anisotropy trajectories are given by

$$\begin{aligned}\Phi^{\text{in}} &:= (\phi^{1,\text{in}}_{\text{dir}}, \phi^{1,\text{in}}_{\text{prop}}, \ldots, \phi^{L,\text{in}}_{\text{dir}}),\\ \Phi^{\text{out}} &:= (\phi^{1,\text{out}}_{\text{dir}}, \phi^{1,\text{out}}_{\text{prop}}, \ldots, \phi^{L,\text{out}}_{\text{dir}}).\end{aligned} \tag{17}$$

Crucially, this *static* deviation from anisotropy in Eq. (17) is complementary to the *dynamic* rotation measures between layers in Eq. (16). We employ embedding rotation $\theta^l_{\text{dir}}$ and the in-distribution alignment $\phi^{l,\text{in}}_{\text{dir}}$ from the direct contribution case as illustrative examples to visualize their relation in Figure 5.

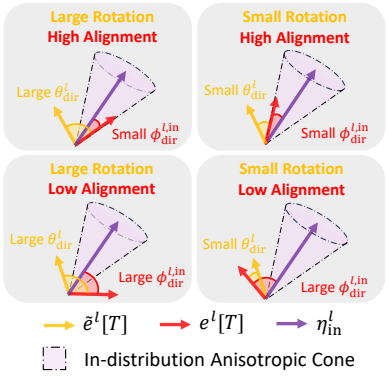



Figure 5: Dynamic rotation and static alignment are complementary but independent characterizations of prediction reliability.

### 4.5 Token-level Confidence Calibration

Previous analyses focused on the first generated token. We now extend this framework to all generated tokens, assigning each token a feature vector $(\Omega, \Theta, \Phi^{\text{in}}, \Phi^{\text{out}}) \in \mathbb{R}^{8L-4}$.

To assess whether the proposed token-level features reflect prediction correctness, we employ the Mahalanobis distance [46]. Specifically, we collect token-level feature vectors from correct predictions and compute their empirical mean and covariance, denoted by $\mu_{\text{corr}}$ and $\Sigma_{\text{corr}}$. Given a test feature vector $\nu$, its Mahalanobis distance to the distribution is $d_{\text{corr}}(\nu) = \sqrt{(\nu - \mu_{\text{corr}})^\top \Sigma^{-1}_{\text{corr}} (\nu - \mu_{\text{corr}})}$. Analogously, we define $d_{\text{inc}}(\nu)$ as the distance to the feature distribution of incorrect predictions.

Tokens from correct predictions should exhibit small $d_{\text{corr}}$ and large $d_{\text{inc}}$, while the opposite is expected for incorrect predictions. We empirically evaluate this separability using Gemma-3-4B-Instruct on Simple-Questions-Wikidata. A reference set is used to estimate $\mu_{\text{corr}}, \mu_{\text{inc}}$ and $\Sigma_{\text{corr}}, \Sigma_{\text{inc}}$. Similarly, we treat correctly predicted samples in the reference set as in-distribution inputs to compute $\tilde{\eta}^{l,\text{in}}$ and $\eta^{l,\text{in}}$, while incorrectly predicted samples are treated as out-of-distribution inputs to compute $\tilde{\eta}^{l,\text{out}}$ and $\eta^{l,\text{out}}$. We combine the token-level features $d_{\text{corr}}$ and $d_{\text{inc}}$ with the uncalibrated confidence to train a calibrator, and aggregate the calibrated confidences to produce response-level uncertainty scores via perplexity in Eq. (2).

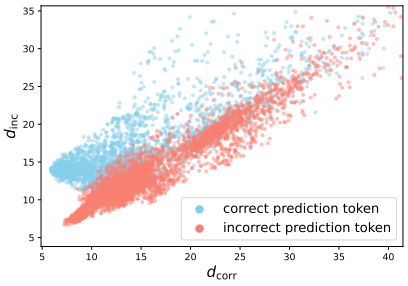



Figure 6: Separation of generated tokens from correct and incorrect predictions using $d_{\text{corr}}$ and $d_{\text{inc}}$.

Algorithm 1 summarizes the complete procedure, including geometry construction, confidence-calibrator training, conformal calibration, and test-time abstention.

**Algorithm 1** Geometry-Calibrated Conformal Abstention

**Require:** Disjoint reference, training, and calibration splits $\mathcal{D}_{\text{ref}}, \mathcal{D}_{\text{train}}, \mathcal{D}_{\text{cal}}$; test input $X_{n+1}$; language model $f$; evaluation function $\xi$; target participation $1-\alpha$.

**Geometry reference construction:** For each $(X, Y) \in \mathcal{D}_{\text{ref}}$, generate its prediction $\widehat{Y} = f(X)$ and determine if the prediction is correct by $J = \xi(X, Y, \widehat{Y})$.

Use predictions with $J = 1$ and $J = 0$ to estimate the expected token embeddings $(\tilde{\eta}^{l,\text{in}}, \eta^{l,\text{in}})$ and $(\tilde{\eta}^{l,\text{out}}, \eta^{l,\text{out}})$ for in- and out-of-distribution queries, respectively.

For each generated token, construct $\nu = (\Omega, \Theta, \Phi^{\text{in}}, \Phi^{\text{out}})$ and estimate $(\mu_{\text{corr}}, \Sigma_{\text{corr}})$ and $(\mu_{\text{inc}}, \Sigma_{\text{inc}})$ from tokens belonging to correct and incorrect predictions, respectively.

**Confidence-calibrator training:** For each $(X, Y) \in \mathcal{D}_{\text{train}}$, generate $\widehat{Y} = f(X)$ and assign it the correctness label $J = \xi(X, Y, \widehat{Y})$.

For each generated token $\widehat{Y}[t]$, construct its geometry feature vector $\nu_t = (\Omega, \Theta, \Phi^{\text{in}}, \Phi^{\text{out}})$ and compute $d_s(\nu_t) = \sqrt{(\nu_t - \mu_s)^\top \Sigma_s^{-1} (\nu_t - \mu_s)}$ for $s \in \{\text{corr}, \text{inc}\}$.

Train the token-level confidence calibrator $r$, using $d_{\text{corr}}(\nu_t)$, $d_{\text{inc}}(\nu_t)$, and raw confidence $p(\widehat{Y}[t] \mid X, \widehat{Y}[<t])$ as input features and the corresponding $J$ of $\widehat{Y}$ as the output label.

**CA calibration:** For $\mathcal{D}_{\text{cal}} = \{(X_i, Y_i)\}_{i=1}^n$, compute $\widehat{Y}_i = f(X_i)$, and $U_i = r(\widehat{Y}_i \mid X_i)$.

Compute the participation threshold $\tau = \text{Quantile}(\frac{\lceil (1-\alpha)(n+1) \rceil}{n}, \{U_i\}_{i=1}^n)$.

Define $\mathcal{K} = \{x : r(f(x) \mid x) \leq \tau\}$.

Determine prediction correctness by $J_i = \xi(X_i, Y_i, \widehat{Y}_i)$.

Let $\mathcal{U} = \{U_i : J_i = 1,\ i = 1, \ldots, n\}$ and $c = |\mathcal{U}|$.

Compute $1 - \beta = \sup\{q \in (0,1) : \text{Quantile}(\frac{\lceil q(c+1) \rceil}{c}, \mathcal{U}) \leq \tau\}$.

Report $\frac{1-\beta}{1-\alpha+1/(n+1)} \frac{c}{n}$ as the plug-in estimate of the conditional-correctness lower bound.

**Test-time decision:** Compute $\widehat{Y}_{n+1} = f(X_{n+1})$ and $U_{n+1} = r(\widehat{Y}_{n+1} \mid X_{n+1})$.

Return $\widehat{Y}_{n+1}$ if $U_{n+1} \leq \tau$, and abstain otherwise.

As shown in Algorithm 1, the reference statistics and confidence calibrator are learned using only the reference and training splits. The resulting uncertainty scoring function $r$ is fixed before it is applied to the calibration and test samples. Conditional on the reference and training splits, the same deterministic mapping $(X_i, Y_i) \mapsto (U_i, J_i)$ is therefore applied to the exchangeable calibration and test examples, preserving the exchangeability required by Lemma 3.2.

## 5 Experiment

**Dataset.** We conduct experiments on six datasets: Natural Questions (NQ) [47] and TruthfulQA (TQA) [48] for open-ended question answering; Simple Questions Wiki (SQW) [40] and SciQ [49] for closed-form question answering; GSM8K [50] and CommonsenseQA (CQA) [51] for reasoning. Each dataset is partitioned into reference, training, calibration, and test splits with a 5:3:1:1 ratio.

**Model Selection.** We use Gemma-3-4B-Instruct [41], LLaMA-3.2-3B-Instruct, and LLaMA-3-8B-Instruct [52] as the base inference models. Prediction correctness is jointly assessed by HHEM-2.1-Open [53], GPT-4o [54], and GPT-5.1 [55], with human validation in Appendix B. The proposed calibrator is trained using the XGBoost library [56]. Our implementation is publicly available.[2]

**Baselines.** The uncertainty scoring function $r$ can be any uncertainty quantification (UQ) method that outputs a scalar reliability estimate. We compare against representative baselines from the widely used llm-polygraph benchmark [57], which span several categories. Likelihood-based methods, such as naïve perplexity [32] and SAR [38], rely on token-level probabilities and entropy of the predictive distribution. Consistency-based approaches estimate uncertainty via response diversity, e.g., Semantic Entropy [33]. Attention-based methods exploit internal attention patterns, including Focus [58] and Attention Score [36]. Embedding-based methods use hidden representations, such as Eigen Score [59] and ATRMD [37]. Finally, verbalized methods are represented by P(True) [60], which directly elicits self-reported confidence from the model. The conformal abstention procedure for each method is performed over 1000 trials, with random resplitting of the calibration and test sets.

### 5.1 Associative Evidence for the Geometry Hypotheses

We examine the associations hypothesized by Hypotheses 1 and 2 using correct and incorrect predictions from the reference sets, with results averaged across datasets. For each sample, we compute the knowledge interaction strength and relative angular distance by layer-wise averaging $\Omega \odot \Theta \in \mathbb{R}^{2L-1}$ and $\Phi^{\text{in}} - \Phi^{\text{out}} \in \mathbb{R}^{2L-1}$, respectively. As shown in Figure 7, correct predictions exhibit stronger knowledge interaction on average, with higher contribution and larger rotation (left), and smaller relative angular distances (right), providing associative evidence consistent with Hypotheses 1 and 2, respectively. Results for these associations are reported separately for each model and dataset in Appendix G.

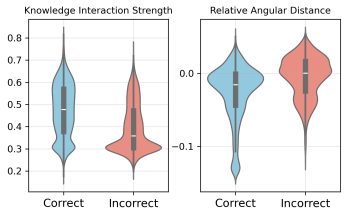



Figure 7: Knowledge interaction strength and relative angular distance between correct and incorrect predictions.

### 5.2 Main Results

The main experimental results using Gemma-3-4B-Instruct as the inference model are shown in Figure 8. Across model participation rate $1 - \alpha$ ranging from 0.1 to 0.9, the proposed method achieves higher empirical conditional correctness, whose corresponding lower bound is given in Eq. (6). This indicates that incorrect predictions receive higher uncertainty scores and are more likely to be abstained from. As participation increases and fewer responses are withheld, conditional correctness approaches the model's base accuracy, reported by dataset in Table 1. Appendix E reports the area under the abstention curve across participation levels.

Table 1: Empirical base accuracy of Gemma-3-4B-Instruct when all responses are retained.

| Dataset | NQ | TQA | SQW | SciQ | GSM8K | CQA |
|---|---|---|---|---|---|---|
| Base accuracy | 0.407 | 0.359 | 0.449 | 0.829 | 0.378 | 0.762 |

In addition, we report comparisons beyond conformal abstention using AUROC and AUPRC on calibration and test data in Table 2. The results follow the same trends as those observed in Figure 8, with the proposed method consistently outperforming the baselines. Experimental results using LLaMA-3.2-3B-Instruct and LLaMA-3-8B-Instruct as the inference models are shown in Appendix C. Across the three inference models, the standard errors of AUROC and AUPRC range from $0.010$ to $0.017$, indicating limited variability across runs. We further estimate the geometry reference statistics and train the confidence calibrator on one dataset, then apply both directly to a different dataset. This cross-dataset setting achieves average AUROC/AUPRC of 0.754/0.782, compared with 0.786/0.813 when training and evaluation use the same dataset, demonstrating good cross-dataset generalizability with only moderate performance degradation (Appendix D). To assess whether the improvement specifically comes from the geometry representation, we compare against confidence-only and dimension-matched PCA controls using the same XGBoost pipeline (Appendix F.1). Separately, we apply XGBoost to the existing uncertainty scores to control for differences in calibration procedures

[2] `https://github.com/rxu0112/Geometry-calibrated-conformal-abstention`

(Appendix F.2). The proposed method remains strongest in both comparisons, indicating that the geometry signals provide information beyond confidence, generic hidden-state features, and learned transformations of the baseline scores. Sensitivity analyses of signal-family weighting, distance metrics, and calibration-set size are reported in Appendix H.

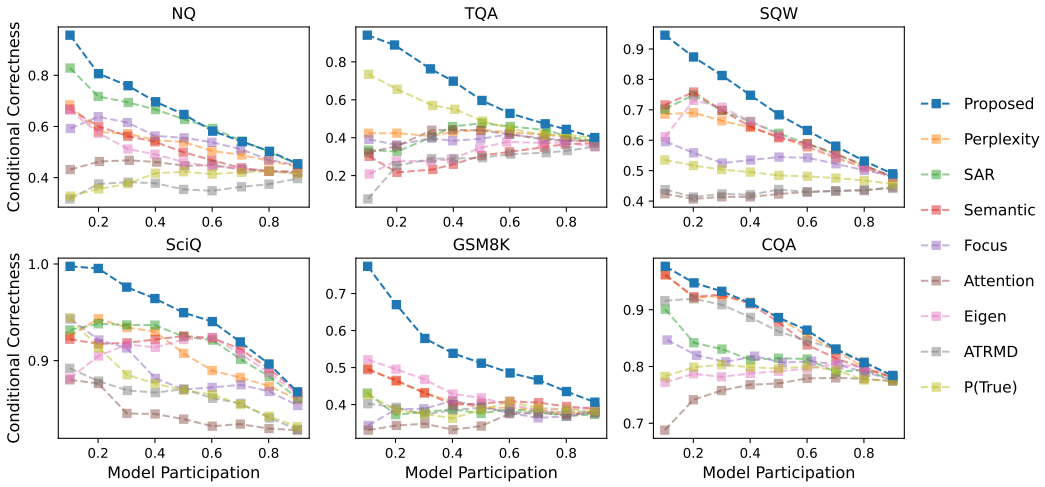



Figure 8: Conditional correctness using different UQ methods as uncertainty scoring functions, with Gemma-3-4B-Instruct as the inference model. We achieve the highest conditional correctness across participation rates.

Table 2: Mean AUROC and AUPRC of different uncertainty scores with Gemma-3-4B-Instruct. Results are averaged over 1000 runs. For each dataset, the highest value is shown in **bold**.

| Method | AUROC | | | | | | AUPRC | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | NQ | TQA | SQW | SciQ | GSM8K | CQA | NQ | TQA | SQW | SciQ | GSM8K | CQA |
| Perplexity | 0.68 | 0.60 | 0.71 | 0.70 | 0.53 | 0.71 | 0.58 | 0.44 | 0.63 | 0.91 | 0.43 | 0.89 |
| SAR | 0.80 | 0.62 | 0.73 | 0.74 | 0.51 | 0.60 | 0.71 | 0.44 | 0.65 | 0.91 | 0.40 | 0.83 |
| Semantic | 0.62 | 0.44 | 0.73 | 0.75 | 0.55 | 0.70 | 0.54 | 0.34 | 0.66 | 0.91 | 0.43 | 0.88 |
| Focus | 0.69 | 0.57 | 0.60 | 0.65 | 0.49 | 0.60 | 0.56 | 0.43 | 0.55 | 0.89 | 0.40 | 0.82 |
| Attention | 0.56 | 0.58 | 0.47 | 0.53 | 0.47 | 0.52 | 0.46 | 0.42 | 0.44 | 0.85 | 0.35 | 0.76 |
| Eigen | 0.59 | 0.48 | 0.72 | 0.73 | 0.56 | 0.58 | 0.52 | 0.34 | 0.62 | 0.90 | 0.44 | 0.79 |
| ATRMD | 0.43 | 0.41 | 0.47 | 0.58 | 0.51 | 0.69 | 0.37 | 0.31 | 0.43 | 0.87 | 0.38 | 0.87 |
| P(True) | 0.50 | 0.70 | 0.56 | 0.60 | 0.52 | 0.57 | 0.40 | 0.62 | 0.50 | 0.88 | 0.39 | 0.80 |
| Proposed | **0.83** | **0.86** | **0.82** | **0.81** | **0.71** | **0.72** | **0.78** | **0.82** | **0.81** | **0.95** | **0.60** | **0.90** |

### 5.3 Permutation-Based Reliance Analysis

We conduct a permutation analysis of the proposed geometry signals $(\Omega, \Theta, \Phi^{\text{in}}, \Phi^{\text{out}})$ by randomly permuting each component during inference and measuring the resulting drops in AUROC and AUPRC. Results are averaged across all inference models and datasets. As shown in Figure 9, permuting each component reduces performance, indicating that the predictor uses information from all four signals. A major limitation of the proposed method is the request for access to internal representations, as discussed in Appendix J.

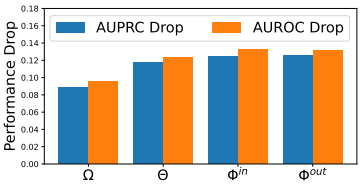



Figure 9: Impact of signal permutation on AUROC and AUPRC.

## 6 Conclusion

In this work, we introduce Conformal Abstention (CA), which enables language models to selectively abstain when their predictions are unreliable, without requiring additional training. We further developed a representation-geometry-based calibration approach that enhances the alignment between prediction confidence and correctness by quantifying how strongly model knowledge contributes to responses, thereby making abstention decisions more effective at filtering out incorrect predictions.

## 7 Acknowledgement

This work was supported in part by the National Key R&D Program of China (Grant No. 2023YFF0725001), the National Natural Science Foundation of China (Grant No. 92370204), the Guangdong Provincial Key Laboratory of Frontier Basic Science for All-domain Intelligence, the Guangdong Basic and Applied Basic Research Foundation (Grant No. 2023B1515120057), the Key-Area Special Project of Guangdong Provincial Ordinary Universities (Grant No. 2024ZDZX1007), the Department of Science and Technology of Guangdong Province (Grant No. 2023CX10X079), the Guangzhou-HKUST(GZ) Joint Funding Program (Grant No. 2023A03J0008), and the Education Bureau of Guangzhou Municipality.

## References

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[2] Kaya Stechly, Karthik Valmeekam, and Subbarao Kambhampati. Chain of thoughtlessness? an analysis of cot in planning. *Advances in Neural Information Processing Systems*, 37:29106–29141, 2024.

[3] Yang Zhang, Hanlei Jin, Dan Meng, Jun Wang, and Jinghua Tan. A comprehensive survey on process-oriented automatic text summarization with exploration of llm-based methods. *arXiv preprint arXiv:2403.02901*, 2024.

[4] Eva Eigner and Thorsten Händler. Determinants of llm-assisted decision-making. *arXiv preprint arXiv:2402.17385*, 2024.

[5] Chandan K Reddy and Parshin Shojaee. Towards scientific discovery with generative ai: Progress, opportunities, and challenges. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 28601–28609, 2025.

[6] Ziqi Yang, Xuhai Xu, Bingsheng Yao, Ethan Rogers, Shao Zhang, Stephen Intille, Nawar Shara, Guodong Gordon Gao, and Dakuo Wang. Talk2care: An llm-based voice assistant for communication between healthcare providers and older adults. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(2):1–35, 2024.

[7] Alexander Tobias Neumann, Yue Yin, Sulayman Sowe, Stefan Decker, and Matthias Jarke. An llm-driven chatbot in higher education for databases and information systems. *IEEE Transactions on Education*, 2024.

[8] Tong Mu, Alec Helyar, Johannes Heidecke, Joshua Achiam, Andrea Vallone, Ian Kivlichan, Molly Lin, Alex Beutel, John Schulman, and Lilian Weng. Rule based rewards for language model safety, 2024.

[9] Bo Wang, Weiyi He, Shenglai Zeng, Zhen Xiang, Yue Xing, Jiliang Tang, and Pengfei He. Unveiling privacy risks in llm agent memory. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25241–25260, 2025.

[10] Bingbing Wen, Jihan Yao, Shangbin Feng, Chenjun Xu, Yulia Tsvetkov, Bill Howe, and Lucy Lu Wang. Know your limits: A survey of abstention in large language models. *Transactions of the Association for Computational Linguistics*, 13:529–556, 2025.

[11] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, 2025.

[12] Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, and Edwin Zhang. Why language models hallucinate, 2025.

[13] Ming Li. Verifiable accuracy and abstention rewards in curriculum rl to alleviate lost-in-conversation, 2025.

[14] Bill Yuchen Lin, Yuntian Deng, Khyathi Chandu, Faeze Brahman, Abhilasha Ravichander, Valentina Pyatkin, Nouha Dziri, Ronan Le Bras, and Yejin Choi. Wildbench: Benchmarking llms with challenging tasks from real users in the wild, 2024.

[15] Anxin Guo and Jingwei Li. Hallucination is a consequence of space-optimality: A rate-distortion theorem for membership testing. *arXiv preprint arXiv:2602.00906*, 2026.

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

[17] Margarida Campos, António Farinhas, Chrysoula Zerva, Mário AT Figueiredo, and André FT Martins. Conformal prediction for natural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 12:1497–1516, 2024.

[18] Anastasios Angelopoulos, Stephen Bates, Jitendra Malik, and Michael I Jordan. Uncertainty sets for image classifiers using conformal prediction. *arXiv preprint arXiv:2009.14193*, 2020.

[19] Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In *Machine learning: ECML 2002: 13th European conference on machine learning Helsinki, Finland, August 19–23, 2002 proceedings 13*, pages 345–356. Springer, 2002.

[20] Bhawesh Kumar, Charlie Lu, Gauri Gupta, Anil Palepu, David Bellamy, Ramesh Raskar, and Andrew Beam. Conformal prediction with large language models for multi-choice question answering. *arXiv preprint arXiv:2305.18404*, 2023.

[21] Allen Z Ren, Anushri Dixit, Alexandra Bodrova, Sumeet Singh, Stephen Tu, Noah Brown, Peng Xu, Leila Takayama, Fei Xia, Jake Varley, et al. Robots that ask for help: Uncertainty alignment for large language model planners. *arXiv preprint arXiv:2307.01928*, 2023.

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

[23] Huanxin Sheng, Xinyi Liu, Hangfeng He, Jieyu Zhao, and Jian Kang. Analyzing uncertainty of llm-as-a-judge: Interval evaluations with conformal prediction, 2025.

[24] Qingni Wang, Tiantian Geng, Zhiyuan Wang, Teng Wang, Bo Fu, and Feng Zheng. Sample then identify: A general framework for risk control and assessment in multimodal large language models. *arXiv preprint arXiv:2410.08174*, 2024.

[25] Zhiyuan Wang, Jinhao Duan, Lu Cheng, Yue Zhang, Qingni Wang, Xiaoshuang Shi, Kaidi Xu, Heng Tao Shen, and Xiaofeng Zhu. Conu: Conformal uncertainty in large language models with correctness coverage guarantees. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6886–6898, 2024.

[26] Jiayuan Su, Jing Luo, Hongwei Wang, and Lu Cheng. Api is enough: Conformal prediction for large language models without logit-access, 2024.

[27] Zhiyuan Wang, Qingni Wang, Yue Zhang, Tianlong Chen, Xiaofeng Zhu, Xiaoshuang Shi, and Kaidi Xu. Sconu: Selective conformal uncertainty in large language models, 2025.

[28] Christopher Mohri and Tatsunori Hashimoto. Language models with conformal factuality guarantees. *arXiv preprint arXiv:2402.10978*, 2024.

[29] 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.

[30] Yasin Abbasi Yadkori, Ilja Kuzborskij, David Stutz, András György, Adam Fisch, Arnaud Doucet, Iuliya Beloshapka, Wei-Hung Weng, Yao-Yuan Yang, Csaba Szepesvári, Ali Taylan Cemgil, and Nenad Tomasev. Mitigating llm hallucinations via conformal abstention, 2024.

[31] Anastasios N Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. *arXiv preprint arXiv:2208.02814*, 2022.

[32] Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. Unsupervised quality estimation for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:539–555, 2020.

[33] Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*, 2023.

[34] Guangya Wan, Yuqi Wu, Jie Chen, and Sheng Li. Reasoning aware self-consistency: Leveraging reasoning paths for efficient LLM sampling. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3613–3635, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics.

[35] Gaurang Sriramanan, Siddhant Bharti, Vinu Sankar Sadasivan, Shoumik Saha, Priyatham Kattakinda, and Soheil Feizi. Llm-check: Investigating detection of hallucinations in large language models. *Advances in Neural Information Processing Systems*, 37:34188–34216, 2024.

[36] Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. Inside: Llms' internal states retain the power of hallucination detection, 2024.

[37] Artem Vazhentsev, Lyudmila Rvanova, Ivan Lazichny, Alexander Panchenko, Maxim Panov, Timothy Baldwin, and Artem Shelmanov. Token-level density-based uncertainty quantification methods for eliciting truthfulness of large language models, 2025.

[38] Jinhao Duan, Hao Cheng, Shiqi Wang, Alex Zavalny, Chenan Wang, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. Shifting attention to relevance: Towards the predictive uncertainty quantification of free-form large language models, 2024.

[39] Angela M Dean and Joseph S Verducci. Linear transformations that preserve majorization, schur concavity, and exchangeability. *Linear algebra and its applications*, 127:121–138, 1990.

[40] Dennis Diefenbach, Thomas Pellissier Tanon, Kamal Deep Singh, and Pierre Maret. Question answering benchmarks for wikidata. In *Proceedings of the ISWC 2017 Posters & Demonstrations and Industry Tracks co-located with 16th International Semantic Web Conference (ISWC 2017), Vienna, Austria, October 23rd - to - 25th, 2017.*, 2017.

[41] Gemma Team. Gemma 3 technical report, 2025.

[42] Yihe Dong, Lorenzo Noci, Mikhail Khodak, and Mufan Li. Attention retrieves, mlp memorizes: Disentangling trainable components in the transformer. *arXiv preprint arXiv:2506.01115*, 2025.

[43] Javier Ferrando, Gerard I Gállego, and Marta R Costa-Jussà. Measuring the mixing of contextual information in the transformer. *arXiv preprint arXiv:2203.04212*, 2022.

[44] Kawin Ethayarajh. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. *arXiv preprint arXiv:1909.00512*, 2019.

[45] Jie Ren, Jiaming Luo, Yao Zhao, Kundan Krishna, Mohammad Saleh, Balaji Lakshminarayanan, and Peter J. Liu. Out-of-distribution detection and selective generation for conditional language models, 2023.

[46] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks, 2018.

[47] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019.

[48] Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods, 2022.

[49] Johannes Welbl, Nelson F. Liu, and Matt Gardner. Crowdsourcing multiple choice science questions, 2017.

[50] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

[51] Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. Commonsenseqa: A question answering challenge targeting commonsense knowledge, 2019.

[52] Aaron Grattafiori et al. The llama 3 herd of models, 2024.

[53] Forrest Bao, Miaoran Li, Rogger Luo, and Ofer Mendelevitch. HHEM-2.1-Open, 2024.

[54] OpenAI. Gpt-4o system card, 2024.

[55] OpenAI. GPT-5.1 Model, 2025.

[56] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794. ACM, August 2016.

[57] Roman Vashurin, Ekaterina Fadeeva, Artem Vazhentsev, Lyudmila Rvanova, Daniil Vasilev, Akim Tsvigun, Sergey Petrakov, Rui Xing, Abdelrahman Sadallah, Kirill Grishchenkov, Alexander Panchenko, Timothy Baldwin, Preslav Nakov, Maxim Panov, and Artem Shelmanov. Benchmarking uncertainty quantification methods for large language models with lm-polygraph. *Transactions of the Association for Computational Linguistics*, 13:220–248, 03 2025.

[58] Tianhang Zhang, Lin Qiu, Qipeng Guo, Cheng Deng, Yue Zhang, Zheng Zhang, Chenghu Zhou, Xinbing Wang, and Luoyi Fu. Enhancing uncertainty-based hallucination detection with stronger focus. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 915–932, Singapore, December 2023. Association for Computational Linguistics.

[59] Gaurang Sriramanan, Siddhant Bharti, Vinu Sankar Sadasivan, Shoumik Saha, Priyatham Kattakinda, and Soheil Feizi. LLM-check: Investigating detection of hallucinations in large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[60] 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.

[61] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert, 2020.

[62] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025.

[63] Xin Qiu and Risto Miikkulainen. Semantic density: Uncertainty quantification for large language models through confidence measurement in semantic space, 2024.

[64] Chenggong Zhang, Haopeng Wang, and Hexi Meng. Hallucination detection and evaluation of large language model, 2026.

[65] Rui Xu, Chao Chen, Yue Sun, Parvathinathan Venkitasubramaniam, and Sihong Xie. Wasserstein-regularized conformal prediction under general distribution shift. In *The Thirteenth International Conference on Learning Representations*, 2025.

# A Additional Theoretical Statement

## A.1 Proof of Lemma 3.2

*Proof.* Consider the deterministic mapping

$$g : ((X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1})) \;\mapsto\; ((U_1, J_1), \ldots, (U_{n+1}, J_{n+1})),$$

where for each $i$,

$$U_i = r(f(X_i) \mid X_i), \quad J_i = \xi(X_i, Y_i, f(X_i)),$$

and $g$ is the composition of the language model $f$, the uncertainty function $r$, and the correctness evaluation $\xi$.

Since $(X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1})$ are exchangeable, for any permutation $\pi$ of $\{1, \ldots, n+1\}$,

$$(X_{\pi(1)}, Y_{\pi(1)}), \ldots, (X_{\pi(n+1)}, Y_{\pi(n+1)}) \overset{d}{=} (X_1, Y_1), \ldots, (X_{n+1}, Y_{n+1}),$$

where $\overset{d}{=}$ indicates distributional equality. Now consider applying $g$ to each of these pairs. Because $g$ acts independently and deterministically on each input pair, we have

$$(U_{\pi(1)}, J_{\pi(1)}), \ldots, (U_{\pi(n+1)}, J_{\pi(n+1)}) = g((X_{\pi(1)}, Y_{\pi(1)}), \ldots, (X_{\pi(n+1)}, Y_{\pi(n+1)})).$$

By Theorem 3.1, if for every permutation $\pi_1$ of the outputs there exists a permutation $\pi_2$ of the inputs such that

$$\pi_1 g((x_1, y_1), \ldots, (x_{n+1}, y_{n+1})) = g(\pi_2(x_1, y_1), \ldots, \pi_2(x_{n+1}, y_{n+1})),$$

then the output sequence is exchangeable. In our case, we can choose $\pi_2 = \pi_1$ because $g$ is applied independently and deterministically to each input. This satisfies the requirement of the theorem for all possible permutations of outputs.

Therefore, the sequence of uncertainty-correctness pairs $(U_1, J_1), \ldots, (U_{n+1}, J_{n+1})$ is exchangeable.

**Remark:** This result holds for any deterministic uncertainty function $r$, correctness evaluation $\xi$, and natural language generation model $f$, provided the underlying prompts and ground truths $(X_i, Y_i)$ are exchangeable. The key intuition is that any permutation of the inputs leads to the same permutation in the outputs, preserving exchangeability under deterministic transformations. □

## A.2 Proof of Lemma 4.2

*Proof.* By definition, the logits are computed as $o = W_E \cdot \varepsilon$, and the ranking $\pi$ is determined by the ordering of the entries:

$$o[\pi(1)] \geq o[\pi(2)] \geq \cdots \geq o[\pi(V)].$$

Now consider a positive scaling of the embedding, $\varepsilon' = \lambda\varepsilon$ with $\lambda > 0$. Then the new logits are

$$o' = W_E \cdot \varepsilon' = W_E \cdot (\lambda\varepsilon) = \lambda(W_E \cdot \varepsilon) = \lambda o.$$

Since $\lambda > 0$, scaling preserves the ordering of the elements of $o$: for any indices $i, j$,

$$o[i] \geq o[j] \implies \lambda o[i] \geq \lambda o[j].$$

Therefore, the ranking induced by $o'$ is identical to that induced by $o$:

$$o'[\pi(1)] \geq o'[\pi(2)] \geq \cdots \geq o'[\pi(V)].$$

This completes the proof. □

### A.3 Proof of Lemma 4.3

*Proof.* We assume the unit token embeddings $\varepsilon, \varepsilon' \in \mathbb{R}^D$ with $\|\varepsilon\|_2 = \|\varepsilon'\|_2 = 1$, and let $\theta$ be the angle between them:

$$\varepsilon^\top \varepsilon' = \cos\theta. \tag{18}$$

Because $\varepsilon$ and $\varepsilon'$ are unit vectors, we have

$$\begin{aligned} \|\varepsilon - \varepsilon'\|_2^2 &= \|\varepsilon\|_2^2 + \|\varepsilon'\|_2^2 - 2\varepsilon^\top \varepsilon' \\ &= 1 + 1 - 2\cos\theta = 2(1 - \cos\theta). \end{aligned} \tag{19}$$

Using the identity $1 - \cos\theta = 2\sin^2(\theta/2)$, it follows that

$$\|\varepsilon - \varepsilon'\|_2 = 2\sin(\theta/2). \tag{20}$$

Let $o = W_E \cdot \varepsilon$ and $o' = W_E \cdot \varepsilon'$ be the corresponding logits. Then

$$o - o' = W_E(\varepsilon - \varepsilon'). \tag{21}$$

Applying the spectral norm of $W_E$, we obtain

$$\|o - o'\|_2 \le \|W_E\|_2 \, \|\varepsilon - \varepsilon'\|_2, \tag{22}$$

and combining with Eq. (20) gives

$$\|o - o'\|_2 \le 2\|W_E\|_2 \, \sin(\theta/2). \tag{23}$$

Define $\gamma := o - o' \in \mathbb{R}^V$ as the change in logits. Let $v_* = \arg\max_v o[v]$ denote the predicted token under $o$. A replacement, i.e., $\arg\max_v o'[v] \neq v_*$, occurs if and only if there exists $v \neq v_*$ such that

$$o'[v] \ge o'[v_*] \iff \gamma[v] - \gamma[v_*] \ge o[v_*] - o[v]. \tag{24}$$

For any $v \neq v_*$, we have

$$|\gamma[v] - \gamma[v_*]| \le |\gamma[v]| + |\gamma[v_*]| \le 2\|\gamma\|_\infty. \tag{25}$$

Therefore, if

$$2\|\gamma\|_\infty \le \min_{v \neq v_*} (o[v_*] - o[v]), \tag{26}$$

then $\gamma[v] - \gamma[v_*] \le o[v_*] - o[v]$ for all $v \neq v_*$, ensuring no replacement occurs.

Since $\|\gamma\|_\infty \le \|\gamma\|_2$ and by Eq. (23) we have $\|\gamma\|_2 \le 2\|W_E\|_2 \, \sin(\theta/2)$, a sufficient condition to avoid replacement is

$$2\|W_E\|_2 \, \sin(\theta/2) \le \min_{v \neq v_*} (o[v_*] - o[v]). \tag{27}$$

This completes the proof. □

## B Correctness Evaluation

To determine whether a generated response is correct, we move beyond token-level overlap metrics (e.g., BERTScore [61]) and instead adopt a systematic, multi-stage semantic evaluation pipeline leveraging HHEM-2.1-Open [53], GPT-4o [54], and GPT-5.1 [55].

First, we convert both the model prediction and the ground-truth answer into natural language statements conditioned on the original question, using Qwen2.5-7B [62]. For example, given the question "Who wrote Pride and Prejudice?", a model prediction such as "It was written by Jane Austen" is reformulated as "The author of Pride and Prejudice is Jane Austen." The ground-truth answer is transformed into a reference statement with the same structure to ensure comparability. We then compute the semantic similarity between the predicted and reference statements using HHEM-2.1-Open, and classify predictions with similarity scores above 0.8 as correct. This formulation evaluates meaning-level agreement rather than surface-level token overlap, which is essential for recognizing semantically equivalent but lexically diverse answers [63].

However, hallucination detection models such as HHEM-2.1-Open evaluate consistency with respect to a reference rather than absolute correctness, and can therefore misclassify semantically valid responses as incorrect when they are not explicitly supported by the reference [64]. To mitigate this issue, we introduce a second-stage verification step. Predictions initially labeled as incorrect are re-evaluated by GPT-4o and GPT-5.1, which act as independent semantic judges. If both models agree that a response is correct, we reassign the sample to the correct class. This conservative consensus mechanism reduces false negatives while maintaining high precision, ensuring that semantically valid answers are not excluded due to limitations of a single evaluator.

**Human Audit and Sensitivity Analysis.** To assess the reliability of the automated correctness evaluator, we randomly sample 50 responses from each of the 18 model-dataset configurations, yielding 900 responses in total. Human annotators label each response as correct or incorrect based on the original question, the generated response, and the reference answer. The human labels agree with the automated evaluator on 836 of the 900 responses, corresponding to an agreement rate of $92.9\%$. To examine the sensitivity of our results to the correctness evaluator, we recompute AUROC and AUPRC on the same 900 responses using either the automated or human correctness labels. The results are reported in Table 3. Replacing the automated labels with human annotations changes AUROC by $0.011$ and AUPRC by $0.033$. These relatively small differences suggest that the main performance conclusions are reasonably stable with respect to the choice of correctness labels.

Table 3: Sensitivity of AUROC and AUPRC to the correctness labels on the human-audited responses.

| Correctness labels | AUROC | AUPRC |
|---|---|---|
| Automated evaluator | 0.841 | 0.877 |
| Human annotation | 0.830 | 0.844 |

## C Additional Experimental Results on the LLaMA Family of Models

Figures 10 and 11 present the conditional correctness–participation trade-off using LLaMA-3.2-3B-Instruct and LLaMA-3-8B-Instruct as the base inference models, respectively. Across participation levels ranging from $1 - \alpha = 0.1$ to $0.9$, our method consistently achieves higher conditional correctness, indicating that the proposed uncertainty scores more effectively distinguish correct from incorrect responses. Consequently, the abstention mechanism selectively filters out less reliable predictions, resulting in improved accuracy among the retained outputs. As participation increases, conditional correctness approaches each model's base accuracy, reported for each dataset in Table 4.

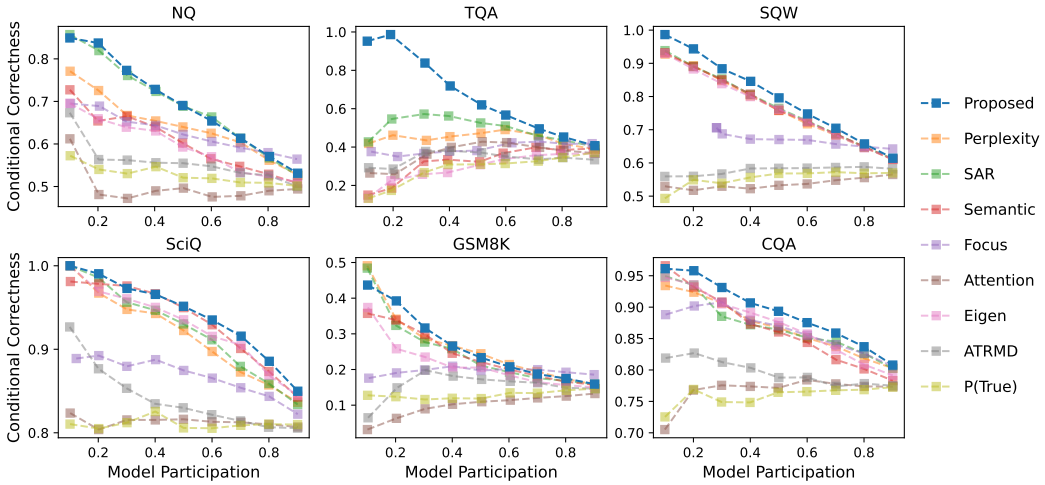



Figure 10: Conditional correctness using different UQ methods as uncertainty scoring functions, with LLaMA-3.2-3B-Instruct as the inference model. We achieve the highest conditional correctness across participation rates.

The experimental results for AUROC and AUPRC using LLaMA-3.2-3B-Instruct and LLaMA-3-8B-Instruct are reported in Table 5 and Table 6, respectively. The proposed method consistently matches

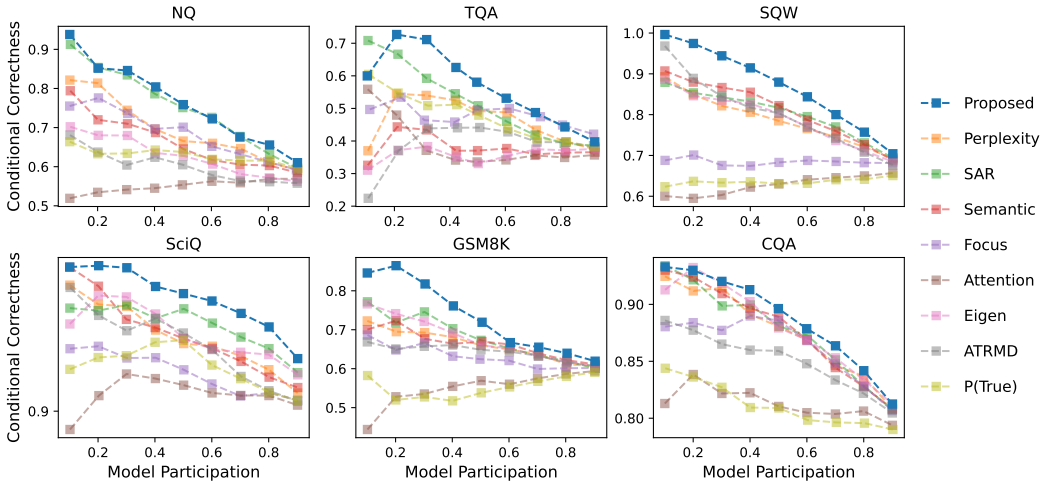



Figure 11: Conditional correctness using different UQ methods as uncertainty scoring functions, with LLaMA-3-8B-Instruct as the inference model. We achieve the highest conditional correctness across participation rates.

Table 4: Empirical base accuracy of the LLaMA inference models when all responses are retained.

| Model | NQ | TQA | SQW | SciQ | GSM8K | CQA |
|---|---|---|---|---|---|---|
| LLaMA-3.2-3B-Instruct | 0.492 | 0.377 | 0.569 | 0.806 | 0.144 | 0.775 |
| LLaMA-3-8B-Instruct | 0.563 | 0.365 | 0.655 | 0.903 | 0.590 | 0.793 |

or outperforms the baselines across both metrics, demonstrating stronger discriminative capability in separating correct from incorrect predictions. This improvement aligns with the gains observed under the conformal abstention framework, further confirming that the proposed uncertainty measure yields a more reliable ranking of predictions.

## D Cross-Dataset Generalization of the Confidence Calibrator

We evaluate whether the proposed geometry-based confidence calibration generalizes across datasets. For each training dataset, we estimate the geometry reference statistics using its reference split and train the confidence calibrator using its training split. Both components are then fixed and applied directly to the test split of each evaluation dataset. Thus, no samples from the evaluation dataset are used to estimate the geometry reference statistics or train the confidence calibrator. We report AUROC and AUPRC because they evaluate the discrimination of the transferred uncertainty score independently of a conformal participation threshold.

Table 7 presents all $6 \times 6$ combinations for the three inference models. Rows denote the evaluation dataset, while columns denote the dataset used to estimate the geometry reference statistics and train the confidence calibrator. Diagonal entries therefore correspond to using the same dataset for training and evaluation, whereas off-diagonal entries represent cross-dataset transfer.

Averaged over the 18 diagonal configurations, using the same dataset for training and evaluation achieves 0.786/0.813 AUROC/AUPRC. Averaged over the 90 off-diagonal configurations, cross-dataset transfer achieves 0.754/0.782, corresponding to decreases of 0.032 and 0.031, respectively. The transferred uncertainty scores therefore retain substantial discrimination, demonstrating good cross-dataset generalizability, although using the same dataset for training and evaluation remains preferable and transfer quality varies across dataset pairs. The diagonal entries are evaluated on the fixed test split and therefore need not coincide exactly with Tables 2, 5, and 6, which use the calibration-and-test evaluation protocol described in Section 5.

This experiment evaluates the discrimination of the transferred uncertainty score rather than the validity of a transferred conformal threshold. A threshold calibrated using one dataset is not guaranteed to preserve the requested participation rate or the conditional-correctness lower bound when applied to a different dataset, as discussed in Appendix J.

Table 5: AUROC and AUPRC of different uncertainty scores with LLaMA-3.2-3B-Instruct. For each dataset, the highest value is shown in **bold**.

| Method | AUROC | | | | | | AUPRC | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | NQ | TQA | SQW | SciQ | GSM8K | CQA | NQ | TQA | SQW | SciQ | GSM8K | CQA |
| Perplexity | 0.71 | 0.64 | 0.76 | 0.75 | 0.77 | 0.69 | 0.67 | 0.51 | 0.81 | 0.92 | 0.40 | 0.87 |
| SAR | **0.77** | 0.70 | 0.77 | 0.76 | 0.73 | 0.69 | 0.74 | 0.52 | 0.81 | 0.93 | 0.39 | 0.88 |
| Semantic | 0.64 | 0.48 | 0.77 | 0.80 | 0.73 | 0.66 | 0.62 | 0.36 | 0.82 | 0.94 | 0.32 | 0.87 |
| Focus | 0.62 | 0.50 | 0.55 | 0.64 | 0.56 | 0.69 | 0.64 | 0.40 | 0.67 | 0.87 | 0.20 | 0.86 |
| Attention | 0.50 | 0.53 | 0.45 | 0.52 | 0.39 | 0.50 | 0.53 | 0.40 | 0.54 | 0.81 | 0.11 | 0.77 |
| Eigen | 0.63 | 0.42 | 0.76 | 0.79 | 0.65 | 0.68 | 0.62 | 0.33 | 0.80 | 0.94 | 0.31 | 0.87 |
| ATRMD | 0.59 | 0.45 | 0.53 | 0.55 | 0.56 | 0.53 | 0.58 | 0.40 | 0.59 | 0.85 | 0.16 | 0.80 |
| P(True) | 0.55 | 0.40 | 0.49 | 0.51 | 0.46 | 0.47 | 0.55 | 0.32 | 0.55 | 0.81 | 0.14 | 0.76 |
| Proposed | **0.77** | **0.88** | **0.81** | **0.82** | **0.78** | **0.74** | **0.75** | **0.87** | **0.86** | **0.95** | **0.42** | **0.90** |

Table 6: AUROC and AUPRC of different uncertainty scores with LLaMA-3-8B-Instruct. For each dataset, the highest value is shown in **bold**.

| Method | AUROC | | | | | | AUPRC | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | NQ | TQA | SQW | SciQ | GSM8K | CQA | NQ | TQA | SQW | SciQ | GSM8K | CQA |
| Perplexity | 0.67 | 0.66 | 0.70 | 0.67 | 0.60 | 0.68 | 0.73 | 0.51 | 0.80 | 0.95 | 0.67 | 0.88 |
| SAR | 0.75 | 0.67 | 0.74 | 0.72 | 0.62 | 0.68 | 0.80 | 0.60 | 0.82 | 0.95 | 0.70 | 0.88 |
| Semantic | 0.63 | 0.52 | 0.74 | 0.67 | 0.61 | 0.68 | 0.68 | 0.39 | 0.83 | 0.95 | 0.66 | 0.88 |
| Focus | 0.64 | 0.66 | 0.49 | 0.59 | 0.52 | 0.67 | 0.69 | 0.52 | 0.69 | 0.92 | 0.63 | 0.86 |
| Attention | 0.49 | 0.50 | 0.47 | 0.53 | 0.46 | 0.54 | 0.57 | 0.45 | 0.63 | 0.91 | 0.55 | 0.81 |
| Eigen | 0.57 | 0.50 | 0.70 | 0.70 | 0.61 | 0.69 | 0.64 | 0.37 | 0.80 | 0.95 | 0.69 | 0.88 |
| ATRMD | 0.54 | 0.58 | 0.71 | 0.64 | 0.58 | 0.63 | 0.62 | 0.41 | 0.83 | 0.94 | 0.64 | 0.85 |
| P(True) | 0.61 | 0.65 | 0.47 | 0.60 | 0.44 | 0.53 | 0.65 | 0.55 | 0.64 | 0.93 | 0.55 | 0.82 |
| Proposed | **0.77** | **0.80** | **0.83** | **0.80** | **0.68** | **0.71** | **0.81** | **0.67** | **0.91** | **0.97** | **0.77** | **0.90** |

# E Area under the Abstention Curve

We report the average area under the abstention curve over 1000 trials, where higher values indicate a greater likelihood of retaining correct predictions, for Gemma-3-4B-Instruct, LLaMA-3.2-3B-Instruct, and LLaMA-3-8B-Instruct in Table 8, Table 9, and Table 10, respectively. The standard deviation is reported after the mean, reflecting the variability across trials. Consistent with the observations in Figures 8, 10, and 11, the proposed method consistently matches or outperforms all baselines across models and datasets, demonstrating stable and robust improvements in abstention performance.

# F Controlled Comparisons for Geometry-Based Confidence Calibration

The proposed method combines geometry-derived information with an XGBoost confidence calibrator. We conduct two complementary comparisons to identify the source of its improvement. First, we hold the XGBoost pipeline fixed and replace the geometry representation with either raw confidence or dimension-matched PCA features. Second, we apply XGBoost to existing uncertainty scores to determine whether their comparison with our method is affected by the absence of learned calibration.

## F.1 Matched Controls for the Geometry Representation

This comparison examines whether the proposed geometry signals provide useful information beyond uncalibrated confidence and generic hidden-state features. The confidence-only control trains the same XGBoost model using only the uncalibrated token confidence. It therefore tests whether learning a nonlinear transformation of confidence alone accounts for the improvement. The PCA control replaces the geometry representation with dimension-matched hidden-state features while retaining the same data splits, distance computation, XGBoost model, and training hyperparameters.

For each generated token, we extract the attention- and MLP-side hidden states from every layer. We retain four principal components from each state in layers $1, \ldots, L-1$ and two components from

Table 7: Cross-dataset generalization of the geometry-based confidence calibrator. Each entry reports AUROC/AUPRC. Rows indicate the evaluation dataset, and columns indicate the dataset used to estimate the geometry reference statistics and train the confidence calibrator. **Bold** diagonal entries use the same dataset for training and evaluation.

**(a) Gemma-3-4B-Instruct**

| Eval./Train | NQ | TQA | SQW | SciQ | GSM8K | CQA |
|---|---|---|---|---|---|---|
| NQ | **0.83 / 0.78** | 0.82 / 0.75 | 0.81 / 0.76 | 0.83 / 0.75 | 0.80 / 0.75 | 0.82 / 0.73 |
| TQA | 0.84 / 0.81 | **0.85 / 0.82** | 0.78 / 0.72 | 0.85 / 0.76 | 0.85 / 0.76 | 0.69 / 0.56 |
| SQW | 0.82 / 0.80 | 0.81 / 0.79 | **0.82 / 0.80** | 0.80 / 0.75 | 0.77 / 0.77 | 0.77 / 0.75 |
| SciQ | 0.75 / 0.93 | 0.72 / 0.92 | 0.71 / 0.92 | **0.80 / 0.95** | 0.73 / 0.93 | 0.77 / 0.94 |
| GSM8K | 0.73 / 0.61 | 0.67 / 0.56 | 0.67 / 0.53 | 0.67 / 0.53 | **0.71 / 0.61** | 0.58 / 0.46 |
| CQA | 0.73 / 0.90 | 0.70 / 0.89 | 0.72 / 0.89 | 0.71 / 0.87 | 0.72 / 0.89 | **0.73 / 0.90** |

**(b) LLaMA-3.2-3B-Instruct**

| Eval./Train | NQ | TQA | SQW | SciQ | GSM8K | CQA |
|---|---|---|---|---|---|---|
| NQ | **0.77 / 0.75** | 0.78 / 0.74 | 0.77 / 0.73 | 0.77 / 0.74 | 0.77 / 0.74 | 0.73 / 0.70 |
| TQA | 0.85 / 0.83 | **0.90 / 0.88** | 0.85 / 0.81 | 0.86 / 0.85 | 0.84 / 0.81 | 0.72 / 0.62 |
| SQW | 0.79 / 0.84 | 0.79 / 0.84 | **0.81 / 0.86** | 0.80 / 0.85 | 0.78 / 0.84 | 0.71 / 0.79 |
| SciQ | 0.80 / 0.94 | 0.79 / 0.94 | 0.82 / 0.95 | **0.83 / 0.95** | 0.79 / 0.94 | 0.69 / 0.91 |
| GSM8K | 0.77 / 0.38 | 0.76 / 0.39 | 0.82 / 0.46 | 0.80 / 0.39 | **0.77 / 0.42** | 0.74 / 0.37 |
| CQA | 0.74 / 0.90 | 0.73 / 0.90 | 0.72 / 0.89 | 0.73 / 0.89 | 0.74 / 0.90 | **0.73 / 0.90** |

**(c) LLaMA-3-8B-Instruct**

| Eval./Train | NQ | TQA | SQW | SciQ | GSM8K | CQA |
|---|---|---|---|---|---|---|
| NQ | **0.77 / 0.81** | 0.76 / 0.79 | 0.76 / 0.77 | 0.74 / 0.77 | 0.77 / 0.81 | 0.77 / 0.81 |
| TQA | 0.82 / 0.69 | **0.81 / 0.67** | 0.82 / 0.70 | 0.79 / 0.69 | 0.79 / 0.68 | 0.78 / 0.67 |
| SQW | 0.80 / 0.89 | 0.74 / 0.86 | **0.83 / 0.91** | 0.81 / 0.89 | 0.72 / 0.85 | 0.74 / 0.85 |
| SciQ | 0.80 / 0.97 | 0.76 / 0.97 | 0.79 / 0.97 | **0.79 / 0.97** | 0.79 / 0.97 | 0.70 / 0.95 |
| GSM8K | 0.69 / 0.76 | 0.67 / 0.74 | 0.66 / 0.74 | 0.64 / 0.71 | **0.68 / 0.77** | 0.61 / 0.68 |
| CQA | 0.66 / 0.88 | 0.66 / 0.88 | 0.66 / 0.88 | 0.68 / 0.88 | 0.62 / 0.85 | **0.71 / 0.89** |

Table 8: Comparison of area under the abstention curve using different uncertainty scores with Gemma-3-4B-Instruct. Standard deviations follow the $\pm$ sign. The highest value is shown in **bold**.

| Method | NQ | TQA | SQW | SciQ | GSM8K | CQA |
|---|---|---|---|---|---|---|
| Perplexity | $0.43 \pm 0.02$ | $0.33 \pm 0.06$ | $0.48 \pm 0.01$ | $0.73 \pm 0.02$ | $0.33 \pm 0.02$ | $0.70 \pm 0.02$ |
| SAR | $0.50 \pm 0.03$ | $0.33 \pm 0.06$ | $0.50 \pm 0.01$ | $0.74 \pm 0.02$ | $0.31 \pm 0.02$ | $0.65 \pm 0.02$ |
| Semantic | $0.41 \pm 0.03$ | $0.24 \pm 0.05$ | $0.50 \pm 0.01$ | $0.73 \pm 0.02$ | $0.33 \pm 0.02$ | $0.70 \pm 0.02$ |
| Focus | $0.44 \pm 0.03$ | $0.31 \pm 0.05$ | $0.43 \pm 0.01$ | $0.71 \pm 0.02$ | $0.31 \pm 0.02$ | $0.64 \pm 0.02$ |
| Attention | $0.36 \pm 0.02$ | $0.32 \pm 0.06$ | $0.34 \pm 0.01$ | $0.68 \pm 0.02$ | $0.28 \pm 0.02$ | $0.61 \pm 0.02$ |
| Eigen | $0.39 \pm 0.02$ | $0.26 \pm 0.05$ | $0.49 \pm 0.01$ | $0.73 \pm 0.02$ | $0.34 \pm 0.02$ | $0.63 \pm 0.02$ |
| ATRMD | $0.29 \pm 0.02$ | $0.23 \pm 0.04$ | $0.34 \pm 0.01$ | $0.69 \pm 0.02$ | $0.31 \pm 0.02$ | $0.69 \pm 0.02$ |
| P(True) | $0.32 \pm 0.02$ | $0.41 \pm 0.06$ | $0.39 \pm 0.01$ | $0.70 \pm 0.02$ | $0.31 \pm 0.02$ | $0.64 \pm 0.02$ |
| Proposed | $\mathbf{0.53 \pm 0.03}$ | $\mathbf{0.51 \pm 0.05}$ | $\mathbf{0.56 \pm 0.01}$ | $\mathbf{0.76 \pm 0.02}$ | $\mathbf{0.43 \pm 0.02}$ | $\mathbf{0.71 \pm 0.02}$ |

each state in the final layer, producing $8L - 4$ PCA features in total. This matches the dimensionality of the proposed geometry representation.

The reference split is used to estimate the correct- and incorrect-prediction distributions in the PCA feature space. On the training split, we compute each token's Mahalanobis distance to both reference distributions and provide these distances, together with the uncalibrated confidence, to XGBoost. At inference, the frozen calibrator produces token-level confidences, which are aggregated into response-level scores using Eq. (2).

As shown in Table 11, the confidence-only control performs substantially worse than the proposed method despite using the same XGBoost model. Thus, the improvement is not attributable merely to learning a transformation of uncalibrated confidence. PCA features improve over confidence alone,

Table 9: Comparison of area under the abstention curve using different uncertainty scores with LLaMA-3.2-3B-Instruct. Standard deviations follow the ± sign. The highest value is shown in **bold**.

| Method | NQ | TQA | SQW | SciQ | GSM8K | CQA |
|---|---|---|---|---|---|---|
| Perplexity | 0.51 ± 0.03 | 0.36 ± 0.06 | 0.61 ± 0.01 | 0.73 ± 0.02 | 0.20 ± 0.02 | 0.70 ± 0.02 |
| SAR | 0.56 ± 0.03 | 0.41 ± 0.06 | 0.62 ± 0.01 | 0.74 ± 0.02 | 0.19 ± 0.02 | 0.70 ± 0.02 |
| Semantic | 0.48 ± 0.03 | 0.26 ± 0.05 | 0.61 ± 0.01 | 0.75 ± 0.02 | 0.19 ± 0.02 | 0.69 ± 0.02 |
| Focus | 0.50 ± 0.03 | 0.31 ± 0.06 | 0.41 ± 0.01 | 0.68 ± 0.02 | 0.16 ± 0.02 | 0.69 ± 0.02 |
| Attention | 0.40 ± 0.03 | 0.30 ± 0.06 | 0.43 ± 0.01 | 0.65 ± 0.02 | 0.08 ± 0.01 | 0.62 ± 0.02 |
| Eigen | 0.48 ± 0.03 | 0.24 ± 0.05 | 0.61 ± 0.01 | 0.74 ± 0.02 | 0.17 ± 0.02 | 0.70 ± 0.02 |
| ATRMD | 0.44 ± 0.03 | 0.28 ± 0.06 | 0.46 ± 0.01 | 0.67 ± 0.02 | 0.13 ± 0.01 | 0.64 ± 0.02 |
| P(True) | 0.42 ± 0.03 | 0.23 ± 0.05 | 0.45 ± 0.01 | 0.65 ± 0.02 | 0.10 ± 0.01 | 0.61 ± 0.02 |
| Proposed | **0.56 ± 0.03** | **0.54 ± 0.06** | **0.64 ± 0.01** | **0.76 ± 0.02** | **0.21 ± 0.02** | **0.72 ± 0.02** |

Table 10: Comparison of area under the abstention curve using different uncertainty scores with LLaMA-3-8B-Instruct. Standard deviations follow the ± sign. The highest value is shown in **bold**.

| Method | NQ | TQA | SQW | SciQ | GSM8K | CQA |
|---|---|---|---|---|---|---|
| Perplexity | 0.56 ± 0.03 | 0.39 ± 0.06 | 0.63 ± 0.01 | 0.76 ± 0.03 | 0.53 ± 0.03 | 0.70 ± 0.02 |
| SAR | 0.60 ± 0.03 | 0.42 ± 0.06 | 0.64 ± 0.01 | 0.76 ± 0.03 | 0.55 ± 0.03 | 0.70 ± 0.02 |
| Semantic | 0.53 ± 0.03 | 0.31 ± 0.06 | 0.65 ± 0.01 | 0.76 ± 0.03 | 0.53 ± 0.03 | 0.70 ± 0.02 |
| Focus | 0.55 ± 0.03 | 0.39 ± 0.06 | 0.55 ± 0.01 | 0.74 ± 0.03 | 0.51 ± 0.03 | 0.69 ± 0.02 |
| Attention | 0.44 ± 0.03 | 0.31 ± 0.06 | 0.50 ± 0.01 | 0.73 ± 0.03 | 0.44 ± 0.02 | 0.65 ± 0.02 |
| Eigen | 0.50 ± 0.03 | 0.30 ± 0.06 | 0.63 ± 0.01 | 0.76 ± 0.03 | 0.54 ± 0.03 | 0.71 ± 0.02 |
| ATRMD | 0.48 ± 0.03 | 0.33 ± 0.05 | 0.64 ± 0.01 | 0.75 ± 0.03 | 0.52 ± 0.03 | 0.68 ± 0.02 |
| P(True) | 0.50 ± 0.03 | 0.39 ± 0.06 | 0.51 ± 0.01 | 0.74 ± 0.03 | 0.44 ± 0.02 | 0.65 ± 0.02 |
| Proposed | **0.61 ± 0.03** | **0.47 ± 0.07** | **0.70 ± 0.01** | **0.78 ± 0.03** | **0.59 ± 0.03** | **0.71 ± 0.03** |

indicating that generic hidden states contain information associated with correctness. However, the geometry representation outperforms PCA for every model, showing that the gain is not explained solely by feature dimensionality or generic hidden-state information.

PCA is unsupervised, whereas the proposed method uses correct and incorrect predictions from the reference split. This comparison therefore controls for feature dimensionality and the downstream distance-and-calibration pipeline, but not for the use of correctness information in constructing the representation.

### F.2 Performance of Calibrated Baselines

The use of a calibrator in our framework raises a natural question: are the observed performance gains driven by the proposed uncertainty signals themselves, or simply by the additional learned calibration step? To disentangle these effects, we apply the same XGBoost calibration procedure to the training-free baselines, whose original formulations produce uncertainty scores without learned calibration. ATRMD is excluded from this comparison because it is itself a training-based method. For each training-free baseline, we report its calibrated output in Table 12, Table 13, and Table 14. We also include the performance of our method in these tables to facilitate direct comparison.

The results indicate that calibration does not necessarily improve these baselines; in fact, their calibrated variants often underperform the raw scores, suggesting susceptibility to overfitting. This behavior is consistent with the nature of these methods. These baselines produce a single scalar uncertainty measure that is already designed to exhibit a monotonic relationship with prediction error. As such, their discriminative capacity is largely intrinsic and requires minimal post-hoc adjustment. Introducing a learned calibrator on top of these one-dimensional signals provides little additional information and can instead amplify noise, likely degrading performance.

In contrast, our method benefits from calibration because it operates on richer, higher-dimensional geometry signals whose relationship to correctness is not trivially monotonic, allowing the calibrator to meaningfully refine the decision boundary.

Table 11: Comparison with matched PCA and confidence-only controls. Values are AUROC/AUPRC averaged across six datasets. Relative reductions are computed against the average performance of the proposed method. The highest performance for each model is shown in **bold**.

| Features | Gemma-3-4B | LLaMA-3.2-3B | LLaMA-3-8B | Average | Relative reduction |
|---|---|---|---|---|---|
| Geometry + confidence | **0.792 / 0.810** | **0.800 / 0.792** | **0.765 / 0.838** | **0.786 / 0.813** | – |
| PCA + confidence | 0.710 / 0.725 | 0.744 / 0.753 | 0.726 / 0.795 | 0.727 / 0.758 | 7.5% / 6.8% |
| Confidence alone | 0.662 / 0.646 | 0.701 / 0.681 | 0.656 / 0.741 | 0.673 / 0.689 | 14.4% / 15.3% |

Table 12: AUROC and AUPRC of calibrated training-free baselines compared with the proposed method, using Gemma-3-4B-Instruct as the base inference model. Arrows next to calibrated baseline results indicate whether performance improves or degrades compared to their original versions, while a dash denotes no change. For each dataset, the highest value is highlighted in **bold**.

| Method | AUROC | | | | | | AUPRC | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | NQ | TQA | SQW | SciQ | GSM8K | CQA | NQ | TQA | SQW | SciQ | GSM8K | CQA |
| Perplexity | 0.67 ↓ | 0.50 ↓ | 0.71 – | 0.71 ↑ | 0.52 ↓ | 0.70 ↓ | 0.54↓ | 0.36↓ | 0.63 – | 0.91 – | 0.40 ↓ | 0.88↓ |
| SAR | 0.79 ↓ | 0.62 – | 0.74 ↑ | 0.74 – | 0.48 ↓ | 0.70 ↑ | 0.66↓ | 0.43↓ | 0.67↑ | 0.92 ↑ | 0.38 ↓ | 0.88↑ |
| Semantic | 0.64 ↑ | 0.41 ↓ | 0.73 – | 0.74 ↓ | 0.51 ↓ | 0.70 – | 0.52↓ | 0.33↓ | 0.67↑ | 0.91 – | 0.38 ↓ | 0.88– |
| Focus | 0.69 – | 0.54 ↓ | 0.66 ↑ | 0.66 ↑ | 0.52 ↑ | 0.59 ↓ | 0.57↑ | 0.37↓ | 0.60↑ | 0.88 ↓ | 0.39 ↓ | 0.80↓ |
| Attention | 0.55 ↓ | 0.70 ↑ | 0.54 ↑ | 0.50 ↓ | 0.54 ↑ | 0.52 – | 0.44↓ | 0.56↑ | 0.49↑ | 0.83 ↓ | 0.40 ↑ | 0.78 ↑ |
| Eigen | 0.57 ↓ | 0.50 ↑ | 0.74 ↑ | 0.74 ↑ | 0.56– | 0.61 ↑ | 0.50↓ | 0.36↑ | 0.70↑ | 0.91 ↑ | 0.44 – | 0.81 ↑ |
| P(True) | 0.51 ↑ | 0.56 ↓ | 0.56 – | 0.50 ↓ | 0.51 ↓ | 0.56 ↓ | 0.43↑ | 0.41↓ | 0.49↓ | 0.83↓ | 0.39 – | 0.79 ↓ |
| Proposed | **0.83** | **0.86** | **0.82** | **0.81** | **0.71** | **0.72** | **0.78** | **0.82** | **0.81** | **0.95** | **0.60** | **0.90** |

## G Geometry-Correctness Associations Across Models and Datasets

Section 5.1 reports the associations between the proposed geometry signals and prediction correctness after averaging across datasets. To examine their consistency across individual settings, we report Cliff's $\delta$ separately for all 18 model-dataset configurations. For each signal, a positive value indicates that a randomly selected correct prediction tends to have a larger value than a randomly selected incorrect prediction, whereas a negative value indicates the reverse.

As shown in Table 15, knowledge interaction has the expected positive direction in 14 of the 18 configurations. The exceptions occur for Gemma-3-4B-Instruct on CQA and for LLaMA-3.2-3B-Instruct on NQ, SQW, and CQA. These results suggest that stronger knowledge interaction is associated with correctness in most settings, but is not sufficient by itself, since the activated knowledge may be irrelevant or misleading.

As shown in Table 16, all 18 configurations have negative Cliff's $\delta$, consistently indicating that correct predictions have smaller relative angular distances. Together, these results provide associative evidence consistent with Hypotheses 1 and 2.

We further evaluate how well each geometry signal individually distinguishes correct from incorrect predictions using AUROC, averaged across the six datasets. As shown in Table 17, relative angular distance provides more consistent discrimination across models, with an overall AUROC of $0.712$, while knowledge interaction is substantially weaker, with an overall AUROC of $0.544$. This supports treating the two signals as complementary: stronger knowledge interaction alone does not ensure correctness because the activated knowledge may be irrelevant or misleading.

## H Sensitivity to Geometry Design and Calibration Size

We examine the sensitivity of the proposed method to two design choices in geometry-based confidence calibration: the weighting of the geometry signals and the distance metric. We also study how the size of the calibration set affects conformal abstention.

Table 13: AUROC and AUPRC of calibrated training-free baselines compared with the proposed method, using LLaMA-3.2-3B-Instruct as the base inference model. Arrows next to calibrated baseline results indicate whether performance improves or degrades compared to their original versions, while a dash denotes no change. For each dataset, the highest value is highlighted in **bold**.

| Method | AUROC | | | | | | AUPRC | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | NQ | TQA | SQW | SciQ | GSM8K | CQA | NQ | TQA | SQW | SciQ | GSM8K | CQA |
| Perplexity | 0.69 ↓ | 0.61 ↓ | 0.77 ↑ | 0.75 – | 0.76↓ | 0.68 ↓ | 0.64 ↓ | 0.44 ↓ | 0.81 – | 0.92 – | 0.37 ↓ | 0.87 – |
| SAR | 0.76 ↓ | 0.68 ↓ | 0.77 – | 0.76 – | 0.72↓ | 0.69 – | 0.75 ↑ | 0.54 ↑ | 0.81 – | 0.93 – | 0.37 ↓ | 0.87↓ |
| Semantic | 0.65 ↑ | 0.72 ↑ | 0.77 – | 0.80 – | 0.73 – | 0.66 – | 0.63 ↑ | 0.53 ↑ | 0.81 ↓ | 0.94 – | 0.30 ↓ | 0.87– |
| Focus | 0.66 ↑ | 0.55 ↑ | 0.66 ↑ | 0.65 ↑ | 0.58↑ | 0.69 – | 0.62 ↓ | 0.42 ↑ | 0.68 ↑ | 0.87 – | 0.17 ↓ | 0.87↑ |
| Attention | 0.51 ↑ | 0.57 ↑ | 0.50 ↑ | 0.52 – | 0.59↑ | 0.53 ↑ | 0.52 ↓ | 0.43 ↑ | 0.57 ↑ | 0.81 – | 0.18 ↑ | 0.79 ↑ |
| Eigen | 0.63 – | 0.49 ↑ | 0.76 – | 0.78↓ | 0.66↑ | 0.68 – | 0.59 ↓ | 0.39 ↑ | 0.81 ↑ | 0.93 ↓ | 0.31 ↓ | 0.87– |
| P(True) | 0.53 ↓ | 0.58 ↑ | 0.50 ↑ | 0.49↓ | 0.50↑ | 0.50 ↑ | 0.53 ↓ | 0.42 ↑ | 0.57 ↑ | 0.80 ↓ | 0.15 ↑ | 0.77 ↑ |
| Proposed | **0.77** | **0.88** | **0.81** | **0.82** | **0.78** | **0.74** | **0.75** | **0.87** | **0.86** | **0.95** | **0.42** | **0.90** |

Table 14: AUROC and AUPRC of calibrated training-free baselines compared with the proposed method, using LLaMA-3-8B-Instruct as the base inference model. Arrows next to calibrated baseline results indicate whether performance improves or degrades compared to their original versions, while a dash denotes no change. For each dataset, the highest value is highlighted in **bold**.

| Method | AUROC | | | | | | AUPRC | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | NQ | TQA | SQW | SciQ | GSM8K | CQA | NQ | TQA | SQW | SciQ | GSM8K | CQA |
| Perplexity | 0.66 ↓ | 0.52 ↓ | 0.70 – | 0.67 – | 0.50 ↓ | 0.68 – | 0.70↓ | 0.48↓ | 0.80 – | 0.94 ↓ | 0.59 ↓ | 0.87↓ |
| SAR | 0.74 ↓ | 0.63 ↓ | 0.74 – | 0.72 – | 0.67 ↑ | 0.68 – | 0.77↓ | 0.51↓ | 0.82 – | 0.95 – | 0.63 ↓ | 0.89↑ |
| Semantic | 0.62 ↓ | 0.44 ↓ | 0.74 – | 0.65 ↓ | 0.59 ↓ | 0.68 – | 0.67↓ | 0.33↓ | 0.83 – | 0.95 – | 0.66 – | 0.89 ↑ |
| Focus | 0.66 ↑ | 0.61 ↓ | 0.57 ↑ | 0.58 ↓ | 0.55 ↑ | 0.66 ↓ | 0.69– | 0.47↓ | 0.70↑ | 0.93 ↑ | 0.62 ↓ | 0.86 – |
| Attention | 0.49 – | 0.46 ↓ | 0.53 ↑ | 0.54 ↑ | 0.53 ↑ | 0.50 ↓ | 0.56↓ | 0.35↓ | 0.67 ↑ | 0.91 – | 0.61 ↑ | 0.79 ↓ |
| Eigen | 0.58 ↑ | 0.51 ↑ | 0.71 ↑ | 0.69 ↓ | 0.60 ↓ | 0.68 ↓ | 0.62↓ | 0.36↓ | 0.81 ↑ | 0.95 – | 0.68 ↓ | 0.89 ↑ |
| P(True) | 0.61 – | 0.50 ↓ | 0.53 ↑ | 0.56 ↓ | 0.52 ↑ | 0.55 ↑ | 0.53↓ | 0.37↓ | 0.68 ↑ | 0.91 ↓ | 0.61 ↑ | 0.82 – |
| Proposed | **0.77** | **0.80** | **0.83** | **0.80** | **0.68** | **0.71** | **0.81** | **0.67** | **0.91** | **0.97** | **0.77** | **0.90** |

### H.1 Sensitivity to Signal-Family Weighting

Our default method concatenates the knowledge contribution, rotation, and anisotropy signals before estimating the means and covariances of correct and incorrect reference predictions. Consequently, deviations in each signal are measured relative to their observed variability, while the covariance captures correlations among signals. The resulting distances to the correct and incorrect reference distributions, together with the raw confidence, are provided to XGBoost.

To test whether the results are affected by differences in scale or dimensionality among the signal families, we separately compute a distance for each family, normalize it by the number of features, and average the three distances with equal weights.

Explicit equal-family weighting performs slightly worse than the default method, as shown in Table 18. This indicates that the improvements are not explained solely by one signal family having more features or a larger numerical scale.

### H.2 Sensitivity to the Distance Metric

The proposed method uses Mahalanobis distance as a covariance-normalized distance and does not convert it into a Gaussian density. We compare it with $k$-NN, shrinkage Mahalanobis, diagonal Mahalanobis, and standardized Euclidean distance. The results are shown in Table 19.

Shrinkage Mahalanobis nearly matches the proposed distance, while the nonparametric $k$-NN alternative remains competitive. The lower performance of diagonal Mahalanobis and standardized Euclidean distance suggests that cross-signal covariance is useful. These results also show that the observed performance does not depend on fitting a Gaussian density.

Table 15: Cliff's $\delta$ for knowledge interaction across model-dataset configurations. Positive values indicate that correct predictions tend to exhibit stronger knowledge interaction than incorrect ones.

| Model | NQ | TQA | SQW | SciQ | GSM8K | CQA |
|---|---|---|---|---|---|---|
| Gemma-3-4B-Instruct | 0.081 | 0.038 | 0.210 | 0.196 | 0.174 | -0.120 |
| LLaMA-3.2-3B-Instruct | -0.177 | 0.012 | -0.197 | 0.016 | 0.134 | -0.069 |
| LLaMA-3-8B-Instruct | 0.035 | 0.150 | 0.429 | 0.451 | 0.065 | 0.141 |

Table 16: Cliff's $\delta$ for relative angular distance across model-dataset configurations. Negative values indicate that correct predictions tend to have smaller relative angular distances than incorrect ones.

| Model | NQ | TQA | SQW | SciQ | GSM8K | CQA |
|---|---|---|---|---|---|---|
| Gemma-3-4B-Instruct | -0.438 | -0.540 | -0.432 | -0.406 | -0.450 | -0.474 |
| LLaMA-3.2-3B-Instruct | -0.514 | -0.296 | -0.438 | -0.417 | -0.269 | -0.461 |
| LLaMA-3-8B-Instruct | -0.389 | -0.407 | -0.488 | -0.448 | -0.297 | -0.461 |

Table 17: AUROC of the individual geometry signals, averaged across the six datasets for each model. Larger knowledge-interaction values and smaller relative-angular-distance values are treated as indicators of correctness.

| Model | Knowledge interaction | Relative angular distance |
|---|---|---|
| Gemma-3-4B-Instruct | 0.548 | 0.728 |
| LLaMA-3.2-3B-Instruct | 0.477 | 0.700 |
| LLaMA-3-8B-Instruct | 0.606 | 0.708 |
| Overall | 0.544 | 0.712 |

### H.3 Sensitivity to Calibration-Set Size

We evaluate calibration-size sensitivity on SQW using its full calibration pool of $n = 1428$ examples. We sample $5\%$, $25\%$, $50\%$, $75\%$, and $100\%$ of this pool, performing 1,000 stratified trials for each size at target participation $1 - \alpha = 0.5$. Here, $c$ denotes the number of correct predictions in the sampled calibration set.

As shown in Table 20, mean participation and conditional correctness remain stable across calibration sizes. Increasing $n$ and $c$ primarily reduces variability and stabilizes the conditional-correctness lower-bound estimate. Because $c$ depends on both the calibration size and the base model's accuracy, more difficult tasks may require larger calibration sets to obtain comparable stability.

## I Memory and Computational Complexity

Let $T$ denote the sequence length, $d$ the hidden dimension, and $L$ the number of Transformer blocks. Since our framework extracts features from both attention and MLP modules, the total number of considered layers is $2L - 1$.

**Memory Complexity.** At each layer $l$, we construct two types of contribution matrices. The MLP contribution forms a diagonal $T \times T$ matrix $C^l_{\mathrm{MLP}}$, which requires $\mathcal{O}(T)$ memory, while the attention contribution $C^l_{\mathrm{Attn}}$ corresponds to a lower-triangular $T \times T$ matrix, requiring $\mathcal{O}(T^2)$ memory. Therefore, the per-layer memory cost is dominated by $\mathcal{O}(T^2)$, and across layers becomes $\mathcal{O}(LT^2)$.

In addition, we only retain hidden states associated with the final input token across layers, resulting in a memory cost of $\mathcal{O}(Ld)$. Other derived quantities, such as embedding rotations and angular features, are also computed on the final token and incur at most $\mathcal{O}(Ld)$ memory.

Combining these terms, the total memory complexity is

$$\mathcal{O}(LT^2 + Ld). \tag{28}$$

Table 18: Comparison of the default method with explicit equal-family weighting. Results are AUROC/AUPRC averaged across six datasets.

| Model | Default | Equal-family weighting |
|---|---|---|
| Gemma-3-4B-Instruct | 0.792 / 0.810 | 0.788 / 0.793 |
| LLaMA-3.2-3B-Instruct | 0.800 / 0.792 | 0.771 / 0.753 |
| LLaMA-3-8B-Instruct | 0.765 / 0.838 | 0.762 / 0.831 |
| Average | 0.786 / 0.813 | 0.774 / 0.792 |

Table 19: Sensitivity to the distance metric. Results are averaged across three models and six datasets.

| Distance metric | AUROC | AUPRC |
|---|---|---|
| Empirical Mahalanobis (proposed) | 0.786 | 0.813 |
| $k$-NN ($k = 10$) | 0.764 | 0.791 |
| Shrinkage Mahalanobis | 0.781 | 0.808 |
| Diagonal Mahalanobis | 0.729 | 0.747 |
| Standardized Euclidean | 0.726 | 0.744 |

**Computational complexity.** The dominant cost arises from computing the proximity function

$$\mathrm{prox}(z, z_i) = \frac{\max\left(0, \|z\|_1 - \|z - z_i\|_1\right)}{\sum_{j=1}^{T} \max\left(0, \|z\|_1 - \|z - z_j\|_1\right)}, \tag{29}$$

where $z, z_i \in \mathbb{R}^d$. Computing $\|z - z_i\|_1$ requires $\mathcal{O}(d)$ operations. For each layer, the attention contribution involves $\mathcal{O}(T^2)$ token pairs, resulting in $\mathcal{O}(T^2 d)$ time. The MLP contribution only involves diagonal terms and is therefore $\mathcal{O}(Td)$, which is negligible compared to the attention term.

Other computations, including embedding rotation and angular distances to in-distribution and out-of-distribution directions, scale as $\mathcal{O}(Td)$ per layer. Therefore, the overall computational complexity is dominated by the pairwise proximity computation:

$$\mathcal{O}(LT^2 d). \tag{30}$$

**Empirical Runtime and Memory.** Overall, the proposed method scales quadratically with sequence length and linearly with the hidden dimension and number of layers. The pairwise proximity computation constitutes the primary computational bottleneck. We complement this asymptotic analysis with measured runtime and peak GPU memory. All methods are evaluated using Gemma-3-4B-Instruct on 50 SciQ examples with the same NVIDIA RTX A6000 GPU.

The proposed method takes approximately $1.0$ second per example, compared with $0.6$ seconds for lightweight scalar scores such as perplexity. It remains faster than AttentionScore, EigenScore, ATRMD, and the methods requiring multiple generations. In particular, SAR and Semantic Entropy are approximately $3.7\times$ and $10\times$ slower, respectively. The proposed method requires $23.4$ GB of peak GPU memory to retain layer-wise hidden states and contribution matrices. This is higher than methods that do not retain internal states, but comparable to the $25.3$ GB required by ATRMD.

## J Limitations

**Access to internal representations.** Although CA can operate with any fixed uncertainty score, our proposed geometry-based score requires access to the language model's internal representations. In particular, it extracts hidden states across layers to compute embedding rotation and alignment. These representations are typically unavailable for black-box or API-only models, limiting the applicability of our geometry-based calibration to open-weight or fully accessible models.

**Reference-set construction.** We estimate the dominant directions by treating correctly predicted samples as in-distribution (ID) references and incorrectly predicted samples as out-of-distribution (OOD) references. This correctness-based proxy does not necessarily reflect genuine distributional membership: incorrect predictions may result from ambiguity or reasoning errors, while correct

Table 20: Sensitivity to calibration-set size on SQW. LLaMA-3.2-3B-Instruct is the base inference model. Values report the mean and standard deviation over 1,000 trials.

| Fraction | $n$ | $c$ | Empirical participation | Empirical conditional correctness | Conditional-correctness lower-bound estimate |
|---|---|---|---|---|---|
| 5% | 71 | 40 | 0.505±0.052 | 0.796±0.027 | 0.761±0.051 |
| 25% | 357 | 203 | 0.500±0.026 | 0.798±0.015 | 0.790±0.022 |
| 50% | 714 | 406 | 0.501±0.020 | 0.797±0.012 | 0.793±0.014 |
| 75% | 1071 | 609 | 0.501±0.018 | 0.796±0.011 | 0.794±0.011 |
| 100% | 1428 | 812 | 0.501±0.016 | 0.796±0.011 | 0.794±0.008 |

Table 21: Runtime and peak GPU memory of different uncertainty scoring methods using Gemma-3-4B-Instruct on 50 SciQ examples. The generation column indicates whether each method requires a single or multiple generations per example.

| Method | Time/50 examples (s) | Time/example (s) | Peak memory (MB) | Generation |
|---|---|---|---|---|
| Proposed | 50 | 1.00 | 23,352 | Single |
| AttentionScore | 60 | 1.20 | 9,012 | Single |
| EigenScore | 90 | 1.80 | 9,014 | Single |
| Focus | 30 | 0.60 | 9,018 | Single |
| Perplexity | 30 | 0.60 | 9,018 | Single |
| P(True) | 30 | 0.60 | 9,024 | Single |
| SAR | 184 | 3.68 | 10,528 | Multiple |
| Semantic Entropy | 500 | 10.00 | 9,068 | Multiple |
| ATRMD | 100 | 2.00 | 25,332 | Single |

predictions may still occur on OOD inputs. Consequently, the estimated dominant directions may be biased, potentially affecting the reliability of the alignment and anisotropy-based measures.

**Calibration-test distribution shift.** Like standard conformal prediction, CA relies on exchangeability between the calibration and test data. We examine violations of this assumption by calibrating on SQW with a target participation rate of $1 - \alpha = 0.5$, and then applying the resulting threshold and SQW-trained confidence calibrator to the other five datasets. Across the 15 model-target-domain configurations, participation gaps range from $-0.0606$ to $0.1643$. Empirical conditional correctness also falls below the source-derived lower-bound estimate in 10 configurations, with a worst-case gap of $-0.5727$. Thus, under distribution shift, neither the requested participation rate nor the source-derived conditional-correctness lower bound is ensured.

As a partial mitigation, we combine CA with a Wasserstein-regularized objective [65] that aligns source- and target-domain score distributions. For SQW-to-NQ transfer using Gemma-3-4B-Instruct, this regularization reduces the absolute participation gap from $0.0556$ to $0.0177$. However, empirical conditional correctness decreases from $0.6520$ to $0.6455$, and its gap from the source-derived lower-bound estimate remains negative ($-0.0302$ versus $-0.0342$). These results suggest that robust score alignment can reduce participation instability, but does not restore conditional correctness. Addressing the latter additionally requires information about target-domain model accuracy and the ranking of correct and incorrect predictions. Combining robust conformal thresholds with lightweight target-domain correctness estimation is therefore an important direction for future work.

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly reflect the paper's contribution.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitation of reliance on internal representation in appendix.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: The abstract and introduction clearly reflect the paper's contribution.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We include all necessary experiment setup information to reproduce the result.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We released the data and code publicly.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide all information, including datasets and models, to understand the results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The standard deviations of results from repeated trials are reported in appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide computation resources and cost in appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research follows the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: The paper focuses on uncertainty quantification for language models. We do not anticipate direct societal impact from the proposed methodology or experiments, as the work is primarily methodological in nature.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite the original papers that produced the models and datasets we used in our work.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: The paper uses language models as part of the core method, including inference generation and the use of hidden states.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.