LLMs Uncertainty Quantification via Adaptive Conformal Semantic Entropy

Hamed Karimi[1] , **Vaishali Meyappan**[1] , **Reza Samavi**[1,2]
[1] Toronto Metropolitan University, Toronto, Ontario, Canada
[2]Vector Institute, Toronto, Ontario, Canada
{hamed.karimi, vaishali.meyappan, samavi}@torontomu.ca

Abstract

LLMs' overconfidence, particularly when hallucinating, poses a significant challenge for the deployment of the models in safety-critical settings and makes a reliable estimation of uncertainty necessary. Existing approaches for uncertainty quantification typically prioritize lexical or probabilistic measures; however, these techniques often ignore the semantic variance of different responses with similar meaning. In this paper, we propose Adaptive Conformal Semantic Entropy (ACSE), a method for estimating prompt-level uncertainty by adaptively measuring semantic dispersion in LLMs outputs. Our uncertainty scoring function is based on clustering semantic entropy of multiple diverse responses to the same prompt. The function adaptively adjusts the uncertainty score based on semantic features of each cluster. To ensure statistical reliability of our score, we use conformal calibration to apply a decision rule to accept/abstain the prompts, providing a finite-sample, distribution-free guarantee such that the error rate among the accepted responses remains bounded by a user-specified tolerance. Our extensive experimental evaluations using different LLMs and datasets, demonstrate that our approach consistently outperforms state-of-the-art uncertainty quantification baselines using discriminative performance, conformal guarantees, and probabilistic calibration indicators. As a highlight, for TriviaQA dataset, AUROC of our approach is 0.88 compared to 0.65 produced by the token entropy approach.

1 Introduction

Despite the wide range application of Large Language Models (LLMs), they remain uncertain in their predictions and often generate incorrect or misleading outputs with high confidence [Shanahan, 2024]. This overconfidence limits their safe deployment in high-stakes domains such as healthcare [Thirunavukarasu *et al.*, 2023], law [Teubner *et al.*, 2023], and scientific research [Liu *et al.*, 2023]. Reliable uncertainty quantification (UQ), defined as the process of estimating the model's confidence in its own predictions, is therefore essential for improving model reliability, safety, and trustworthiness [Chang *et al.*, 2024].

Related Work. Most existing methods for uncertainty estimation in LLMs rely on token-level signals, such as entropy over next-token distributions [Duan *et al.*, 2023; Fadeeva *et al.*, 2024; Yarie *et al.*, 2024] or average sequence log-likelihood [Yao *et al.*, 2019; Huang *et al.*, 2023]. While computationally efficient, these approaches capture only surface-level lexical variability rather than uncertainty over meaning [Shorinwa *et al.*, 2025]. In autoregressive generation, LLMs often assign high probability to multiple near-synonymous tokens, producing outputs that are probabilistically coherent yet semantically ambiguous or incorrect. As a result, token-level entropy cannot distinguish superficial phrasing variation from genuine semantic ambiguity. In particular, entropy may remain low even when probability mass is spread across responses with mutually inconsistent meanings, creating a misleading impression of confidence [Brown *et al.*, 2020; Holtzman *et al.*, 2020].

More recent work has shifted toward semantic-level uncertainty estimation, notably Semantic Entropy (SE) [Kuhn *et al.*, 2023], which aggregates uncertainty by clustering semantically equivalent responses using Natural Language Inference. However, SE relies on hard cluster assignments, forcing each response into a single cluster and ignoring overlapping semantic regions, which can yield unstable uncertainty estimates. To improve reliability, conformalized approaches [Kaur *et al.*, 2024] employ dynamic clustering with finite-sample guarantees, while the Conformalized Abstention Policy [Tayebati *et al.*, 2025] integrates reinforcement learning to learn instance-specific risk thresholds. Despite their guarantees, these methods treat uncertainty as a scalar signal and remain insensitive to cluster brittleness. As a result, they are vulnerable to the wrong-consensus trap, in which lexically consistent but factually incorrect response clusters satisfy conformal thresholds, masking underlying semantic error.

Proposal. To address these limitations, we propose *Adaptive Conformal Semantic Entropy* (ACSE), a model-independent method that quantifies uncertainty in pretrained LLMs based on dispersion in meaning. Rather than relying on token-level signals, ACSE estimates prompt-level uncertainty in a continuous semantic vector space, capturing dispersion in response meaning. For a finite set of representative prompts (e.g.,

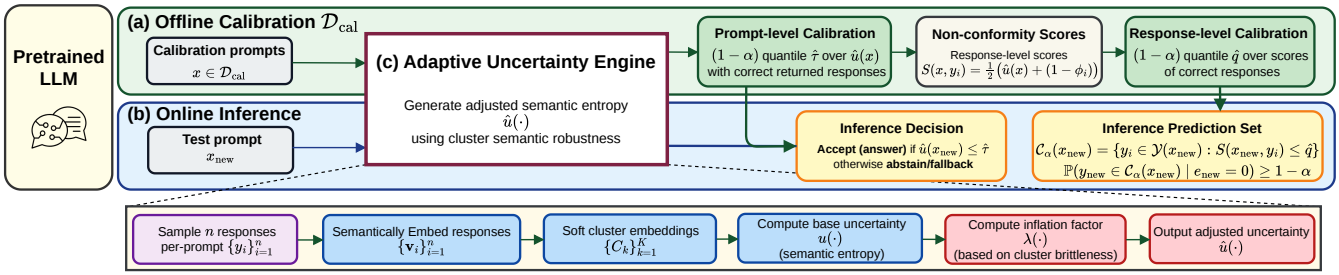


Figure 1: ACSE Pipeline. (a) To calibrate a pretrained LLM, for each prompt $x \in \mathcal{D}_{cal}$, we use an Adaptive Uncertainty Engine (AUE) to compute an adjusted uncertainty score, $\hat{u}(x)$. The $\hat{u}(x)$ is then used to compute a cutoff for a conformalized decision rule (prompt-level) and prediction set in response-level for the user defined upper bound error guarantee of α. (b) In inference time, for a new prompt, the same AUE is used to compute $\hat{u}(x)$ to compare with the calibration cutoff. (c) AUE starts with sampling n responses, embeds them, and soft-clusters the embeddings, leading to the computation of base semantic entropy $u(x)$.

1,000 calibration prompts), we sample multiple responses per prompt, embed each response using a sentence encoder whose cosine similarity reflects semantic proximity [Reimers and Gurevych, 2019], and cluster the embeddings into coherent semantic groups. We then estimate probability mass for each group via soft assignments and derive a base uncertainty score from the normalized entropy of the resulting distribution. This base semantic uncertainty reflects the degree of meaning dispersion: higher dispersion–corresponding to multiple semantically distinct responses–indicates greater uncertainty, whereas semantically consistent responses imply lower uncertainty [Yadkori *et al.*, 2024]. We subsequently adjust the uncertainty score when cluster structure indicates brittleness, such as weak support for the dominant cluster, high internal diversity, or other risk patterns that empirically correlate with errors under distribution shift. Finally, we calibrate the adjusted uncertainty score using conformal prediction on held-out calibration prompts [Vovk *et al.*, 2005] to learn a cutoff that serves two purposes: (1) making prompt-level answer/abstain decisions and (2) generating response-level prediction sets to support user decision-making. Figure 1 provides an overview of the approach, which is illustrated by a toy example.

Toy Example. Consider a user querying a pretrained geographical LLM with the prompt *What is the capital of Australia?*. The model returns *Sydney* with probability 70% (high confidence) despite its semantic incorrectness, and *Canberra* with probability 30% (low confidence). Selecting the most likely sequence yields a *wrong consensus*, a common failure mode in open-ended generation. Assume the user has calibrated the model on a held-out set of geographical prompts and identified 0.58 as a conformity cutoff corresponding to 90% confidence under the assumption that new prompts are drawn from the same distribution [Vovk *et al.*, 2005]. For a new prompt, ACSE samples 10 responses using fixed decoding settings: 9 contain *Sydney* and 1 contains *Canberra*. These responses are embedded, clustered by meaning, and assigned soft semantic memberships via normalized cosine similarity to cluster centroids. Since assignments are probabilistic, clusters are non-disjoint. Averaging memberships across responses yields mean cluster weights of 0.87 for *Sydney* and 0.13 for *Canberra*. The resulting base semantic uncertainty score, given by the normalized entropy of these weights, is 0.55. Because this score falls below the abstention threshold ($0.55 < 0.58$), the model would return the dominant *Sydney* cluster. However, semantic entropy alone can be overly optimistic when clusters are internally brittle, e.g., when cluster membership is unstable or weakly representative. Our approach accounts for such brittleness by adjusting the uncertainty score. In this example, incorporating cluster quality factors (Section 3.1) increases the score to 0.62. Since $0.62 > 0.58$, the model abstains from returning the incorrect answer. While sampling and post-processing incur additional cost, this overhead is justified in safety-critical applications such as using LLMs for mental health support.

Contributions. We make the following contributions. First, we propose ACSE, a method for estimating entropy-based uncertainty at the semantic level for a given prompt. Second, we introduce an adjusted uncertainty score that leverages cluster robustness features to adaptively inflate semantic uncertainty, explicitly penalizing ambiguous response semantics and mitigating hallucinations. Third, we incorporate a post-hoc conformal calibration phase on held-out prompts to learn a cutoff over adjusted uncertainty estimates aligned with response correctness and with formal guarantee, yielding decision rules in both prompt-level and response level. Finally, we demonstrate through extensive experiments on open-domain question answering benchmarks that our adaptive semantic-level uncertainty modeling achieves superior discrimination and calibration compared to state-of-the-art baselines. The extended version of this paper can be found at *https://arxiv.org/abs/2605.04295*, which includes additional experimental evaluations, the computation overhead analysis, and proofs of the theorems.

2 Prompt-Level Uncertainty Estimation

ACSE calibrates a pretrained LLM to a new domain using a finite set of labeled calibration prompts $\mathcal{D}_{\text{cal}}$ of size M. Calibration begins by sampling multiple responses per prompt to assess the consistency of the model's generated meanings.

2.1 Response Sample Generation

Let $\mathcal{V}$ denote the token vocabulary. A generated response is a finite token sequence $y_i = \langle y_{i,1}, \ldots, y_{i,T_i} \rangle$ terminated by an EOS token in T_i decoding steps. For a fixed input

prompt $x \in \mathcal{X}$, we draw a set of n independent responses $\mathcal{Y}(x) = \{y_1, \ldots, y_n\}$ from a language model with parameters θ. At each decoding step t, the model defines a next-token probability distribution over tokens $v \in \mathcal{V}$ as,

$$\pi_{i,t}(v) = p_\theta(y_{i,t} = v \mid x, y_{i,1:t-1}), \tag{1}$$

where $\sum_{v \in \mathcal{V}} \pi_{i,t}(v) = 1$ and $y_{i,t}$ denotes the t-th token of response y_i. To balance diversity and semantic coherence, we adopt nucleus (top-η) sampling [Holtzman *et al.*, 2020], which dynamically truncates the distribution at each step. Specifically, the nucleus $\Gamma_\eta(x, i, t)$ is the smallest subset of tokens whose cumulative probability exceeds a threshold $\eta \in \mathbb{R}^{(0,1]}$,

$$\Gamma_\eta(x, i, t) = \arg\min_{L \subseteq \mathcal{V}} \left\{ |L| : \sum_{v \in S} \pi_{i,t}(v) \geq \eta \right\}. \tag{2}$$

The truncated distribution is obtained by renormalizing over Γ_η,

$$\pi^{(\eta)}_{i,t}(v) = \begin{cases} \dfrac{\pi_{i,t}(v)}{\sum_{v \in \Gamma_\eta(x,i,t)} \pi_{i,t}(v)}, & v \in \Gamma_\eta(x, i, t), \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

Tokens are sampled as $y_{i,t} \sim \pi^{(\eta)}_{i,t}(\cdot)$ and appended until EOS or a length limit is reached. Repeating this procedure yields the response set $\mathcal{Y}(x)$. The variability of $\mathcal{Y}(x)$ provides an empirical proxy for model confidence. When the model is confident, stochastic decoding produces lexically diverse but semantically equivalent responses; when uncertain, samples diverge into multiple, often conflicting, meanings. The sample size n controls a trade-off between computational cost and resolution: larger n captures finer-grained semantic variation, while smaller n suffices when responses concentrate on a single meaning. In the next step, we quantify this dispersion by embedding $\mathcal{Y}(x)$ into a semantic space where cosine similarity reflects semantic proximity.

2.2 Semantic Embedding

We represent each sampled response by its underlying meaning rather than its surface form, so that paraphrases are embedded nearby while semantically distinct answers are well separated. This representation enables reliable grouping of responses into semantic clusters for downstream uncertainty estimation. Each generated response $y_i \in \mathcal{Y}(x)$ is mapped into a continuous semantic space using a pre-trained sentence encoder $f : \mathcal{Y} \to \mathbb{R}^d$, producing embeddings $\mathcal{E} = \{\mathbf{v}_i = f(y_i) \in \mathbb{R}^d \mid y_i \in \mathcal{Y}(x)\}$. We ℓ_2-normalize these embeddings such that $\|\mathbf{v}_i\|_2 = 1$ and measure similarity using cosine geometry [Cer *et al.*, 2018; Reimers and Gurevych, 2019; Gao *et al.*, 2021]. This choice reflects semantic relatedness across lexical and syntactic variation, provides bounded and scale-invariant similarity scores, and yields more stable neighborhoods in high-dimensional spaces than Euclidean distance. The resulting embeddings capture the latent semantic structure of model outputs and form the basis for measuring dispersion in meaning. The effectiveness of this representation depends on selecting a sentence encoder that reliably separates semantic similarity from surface variation.

Using the embeddings $\mathcal{E}$, we define a binary error function $e : \mathcal{X} \times \mathcal{Y} \to \{0, 1\}$ that assigns error labels to prompt-response pairs where $e(x, y_i) = \mathbb{1}\{y_i \text{ is incorrect}\}$. Labels are determined automatically via semantic matching [Cer *et al.*, 2018; Reimers and Gurevych, 2019]. Given a reference response $\hat{y}$ for prompt x and a cosine similarity threshold $\tau_{\cos} \in (0, 1)$, a response y_i is labeled correct if $\cos(\mathbf{v}_{y_i}, \mathbf{v}_{\hat{y}}) \geq \tau_{\cos}$. In the following, we leverage semantic clusters of $\mathcal{E}$ to derive soft assignments that preserve the semantic ranking induced by the encoder.

2.3 Soft Clustering of Semantic Embedding

To transform the unstructured distribution of response embeddings into a meaningful uncertainty signal, we aggregate semantically proximal embeddings into clusters that represent distinct meanings. This serves two purposes: (1) collapsing lexically diverse but semantically equivalent responses into a single equivalence class, and (2) exposing higher-level semantic structure whose relative prevalence reflects model uncertainty. We apply Hierarchical Agglomerative Clustering (HAC) [Murtagh and Contreras, 2017] to the embedding set $\mathcal{E}$ under cosine geometry. Using unit-normalized embeddings, cosine dissimilarity reduces to

$$\operatorname{dist}(\mathbf{v}, \hat{\mathbf{v}}) = 1 - \mathbf{v}^\top \hat{\mathbf{v}}, \quad \text{s.t.} \quad \|\mathbf{v}\|_2 = \|\hat{\mathbf{v}}\|_2 = 1. \tag{4}$$

HAC constructs a dendrogram by iteratively merging clusters with minimum average pairwise dissimilarity (average linkage), balancing sensitivity to local neighborhoods and global coherence. Cutting the dendrogram at a threshold $\epsilon \in \mathbb{R}^{(0,1)}$ yields a set of K semantically coherent clusters

$$\mathcal{J} = \{C_1, \ldots, C_K\}, \qquad \text{where} \qquad \bigcup_{k=1}^{K} C_k = \mathcal{E}. \tag{5}$$

Lower ϵ enforces strict semantic agreement and higher ϵ allows broader conceptual grouping. This adaptive thresholding aligns cluster granularity with the intrinsic density of the embedding space.

To avoid overstating certainty near cluster boundaries, we adopt soft cluster assignments. Each cluster C_k is represented by a centroid $\mathbf{c}_k = \sum_{\mathbf{v} \in C_k} \mathbf{v} / |C_k|$. For a response embedding $\mathbf{v}_i$, we compute its similarity to each centroid as,

$$a_{ik} = \frac{1}{2}\Big(1 + \cos(\mathbf{v}_i, \mathbf{c}_k)\Big) \in \mathbb{R}^{[0,1]}, \tag{6}$$

which is a parameter-free, monotonic transformation that preserves angular geometry without temperature tuning. Then, normalizing across clusters yields soft assignments

$$s_{ik} = \frac{a_{ik}}{\sum_{k=1}^{K} a_{ik}} \qquad \text{s.t.} \qquad \sum_{k=1}^{K} s_{ik} = 1, \tag{7}$$

where s_{ik} quantifies the degree to which response y_i supports semantic cluster C_k. Assignments concentrate when a response aligns strongly with one cluster, and disperse when it lies between competing meanings, preserving signals of semantic ambiguity. Finally, we aggregate soft assignments across the n sampled responses for prompt x to

obtain a prompt-level distribution over meanings $\mathcal{P}(C_k) = n^{-1}\sum_{i=1}^{n} s_{ik}$ where $\sum_{k=1}^{K}\mathcal{P}(C_k) = 1$. We define the *Semantic Entropy (SE)* as,

$$\mathcal{H}_{\text{sem}}(x) = -\sum_{k=1}^{K}\mathcal{P}(C_k)\log\mathcal{P}(C_k)\,, \tag{8}$$

and normalize it to ensure comparability across prompts with different K,

$$u(x) = \begin{cases} \dfrac{\mathcal{H}_{\text{sem}}(x)}{\log K}\,, & K \geq 2\,, \\ 0\,, & K = 1\,. \end{cases} \tag{9}$$

The normalized score $u(x) \in \mathbb{R}^{[0,1]}$ quantifies semantic dispersion: low values indicate strong agreement on a single meaning, while high values reflect uncertainty arising from competing semantic interpretations.

3 Conformalized Adaptive Semantic Entropy

Conformal Prediction. Conformal Prediction (CP) is a distribution-free framework for calibrating set-valued predictions with finite-sample guarantees. Given a calibration dataset $\mathcal{D}_{\text{cal}} = \{(x_i, y_i)\}_{i=1}^{M}$ and a test point (x_{M+1}, y_{M+1}) drawn from the same distribution, CP ensures

$$\mathbb{P}(y_{M+1} \in \mathcal{C}(x_{M+1})) \geq 1 - \alpha\,, \tag{10}$$

where $\alpha \in \mathbb{R}^{(0,1)}$ is the miscoverage level and $\mathcal{C}$ is the prediction set [Vovk *et al.*, 2005]. CP relies on a non-conformity score $S : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^{+}$ that measures how poorly a candidate output conforms to the calibration data. For confidence level $1-\alpha$, the prediction set is

$$\mathcal{C}(x) = \{\hat{y} \in \mathcal{Y} :\ S(x, \hat{y}) \leq \tau_\alpha\}\,, \tag{11}$$

where τ_α is the empirical $(1-\alpha)$-quantile of calibration scores. This guarantee is model- and distribution-agnostic with exchangeability assumption, i.e., the joint distribution is invariant under permutations of the calibration samples. We refer to [Angelopoulos and Bates, 2023] for a comprehensive treatment. We adapt CP calibration to LLMs for two complementary purposes: (1) to derive a prompt-level decision rule that determines whether to answer or abstain, and (2) to construct a response-level prediction set over sampled outputs that supports user decision-making by identifying responses satisfying conformal coverage. Achieving these goals hinges on a critical component of CP: the fidelity of the non-conformity score. Since the quality of conformal prediction sets depends directly on this score, we introduce an adjusted uncertainty score tailored to LLMs (Section 3.1).

3.1 Adjusted Semantic Uncertainty

Although the base semantic uncertainty $u(x)$ in (9) captures dispersion across meanings, it does not account for structural properties of semantic clusters that correlate with prediction difficulty. In particular, $u(x)$ can underestimate uncertainty in low-sample regimes, semantically heterogeneous clusters with weak separation, and sparsely supported clusters that are vulnerable to distribution shift. To address this, we amplify $u(x)$ using the *odds of uncertainty*, $O(u, x) = u(x)/(1-u(x))$, which represents the uncertainty-confidence ratio. To capture cluster brittleness, we define a prompt-level inflation factor $\lambda : \mathcal{X} \to \mathbb{R}^{[1,\lambda_{\max}]}$ to scale these odds such that $O(\hat{u}, x) = \lambda(x)\cdot O(u, x)$, leading to the adjusted uncertainty

$$\hat{u}(x) = \frac{\lambda(x)\,u(x)}{1 + \big(\lambda(x) - 1\big)\,u(x)} \in \mathbb{R}^{[u(x),\,1]}\,. \tag{12}$$

This bounded, monotone, and order-preserving transform ensures $\hat{u}(x) \geq u(x)$ with equality iff $\lambda(x) = 1$ or $u(x) \in \{0, 1\}$. Inflation increases with both u and λ, peaking at intermediate $u(x)$ values to prioritize sensitive borderline decisions. Unlike response-level error labels $e(x, y)$, the inflation factor $\lambda(x)$ operates strictly at the prompt level by aggregating properties of the full response set $\mathcal{Y}(x)$.

To make $\lambda(x)$ interpretable and data-driven, we synthesize five normalized, prompt-level robustness features from the embedding geometry using ℓ_2-normalized response embeddings $\{\mathbf{v}_i\}_{i=1}^{n}$, soft assignments s_{ik} of response y_i to cluster C_k, and the agglomerative cluster set $\mathcal{J}$ with unit-norm centroids $\{\mathbf{c}_k\}_{k=1}^{K}$, where $k^\star = \arg\max_k \mathcal{P}(C_k)$ denotes the dominant cluster and $i^\star = \arg\max_i s_{ik^\star}$ denotes the response most representative of the dominant cluster $C_{k^\star}$.

(1) Semantic Entropy. The normalized semantic entropy $u(x)$ measures multimodality across clusters and serves as the primary ambiguity signal. Larger $u(x)$ directly increases $\lambda(x)$, inflating uncertainty when probability mass is spread across competing meanings.

(2) Centroid Distance. We assess if the dominant response is geometrically supported by its cluster using the centroid similarity $a_{i^\star k^\star}$ in (6). The distance $\tilde{a}(x) = 1 - a_{i^\star k^\star}$ is small when the representative response lies near the cluster center and large when it is atypical, indicating structural instability. We set $\tilde{a}(x) \propto \lambda(x)$ so that uncertainty increases whenever the dominant prediction lacks sufficient geometric support.

(3) Dominant Cluster Dispersion. To quantify internal coherence, we define the intra-cluster dispersion

$$d_{k^\star}(x) = \frac{1}{2|C_{k^\star}|}\sum_{\mathbf{v}\in C_{k^\star}}\Big(1 - \cos\big(\mathbf{v},\,\mathbf{c}_{k^\star}\big)\Big) \in \mathbb{R}^{[0,1]}\,. \tag{13}$$

Low values indicate tight semantic agreement, while high values reflect weak internal consistency; $d_{k^\star}(x)$ increases $\lambda(x)$ proportionally; thus, uncertainty is inflated whenever the dominant cluster lacks internal coherence.

(4) Dominant Cluster Size. To penalize fragile consensus supported by few samples, we define the sparsity feature

$$g_{k^\star}(x) = \min\Big\{1,\ \frac{\kappa}{|C_{k^\star}|}\Big\} \in \mathbb{R}^{[0,1]}\,, \tag{14}$$

where $\kappa = \text{median}_{x\in\mathcal{D}_{\text{cal}}}\big(\max_k |C_k|\big)$ is a dataset-specific scaling constant computed once on $\mathcal{D}_{\text{cal}}$. Smaller dominant clusters yield larger $g_{k^\star}(x)$ directly proportional to $\lambda(x)$ to inflate uncertainty under sparse semantic support.

(5) Margin to Threshold (Overconfidence Suppression). To suppress unwarranted confidence in the low-uncertainty regime, we introduce the margin-to-threshold feature $m(x)$. Using the calibration set $\mathcal{D}_{\text{cal}}$, we define a label-free reference threshold $\tau_{\text{ref}} = \text{Quantile}_{\gamma}(\{u(x) : x \in \mathcal{D}_{\text{cal}}\})$ where $\gamma \geq 0.5$ is user specified. This threshold is fixed after calibration and reused during deployment to compute $\lambda(x)$, thereby preserving the validity of subsequent conformal calibration. We define the normalized margin

$$m(x) = \max\Big\{0,\ 1 - \frac{u(x)}{\tau_{\text{ref}}}\Big\} \ \in \mathbb{R}^{[0,1]}\,, \tag{15}$$

which is large when $u(x) \ll \tau_{\text{ref}}$, indicating extreme, and potentially miscalibrated confidence. We incorporate $m(x)$ into $\lambda(x)$ so that overconfident prompts receive additional inflation, nudging borderline cases toward abstention, while the $\max$ operator ensures the penalty vanishes whenever $u(x) \geq \tau_{\text{ref}}$.

All features are computed identically during calibration and inference to ensure representational consistency. The constants κ and τ_{ref} are derived once from the unlabeled calibration set $\mathcal{D}_{\text{cal}}$ and frozen thereafter, preserving conformal validity. To design our monotone inflation function λ, we introduce a non-negative weight vector $\mathbf{w} = \langle w_u, w_{\tilde{a}}, w_d, w_g, w_m \rangle \in [0,1]^5$ that encodes the relative contribution of each structural semantic feature of the feature set $\mathcal{F} = \{u, \tilde{a}, d_{k^\star}, g_{k^\star}, m\}$ where $\sum_{l \in \mathcal{F}} w_l = 1$. These assigned weights are a design choice and not learned parameters, accounting for prioritization over cluster brittleness features. We aggregate the feature set $\mathcal{F}$ by weighted average, into the normalized composite brittleness metric $B(x) = \sum_{l \in \mathcal{F}} w_l\, l(x) \ \in \mathbb{R}^{[0,1]}$ where larger $B(x)$ indicates a less reliable prompt due to cluster brittleness. We then define the inflation mapping $\lambda(x) = 2/(2 - B(x)) \in \mathbb{R}^{[1,2]}$, which satisfies $\lambda_{\min}(x) = 1$ for benign prompts ($B(x) = 0$) and approaches the bounded maximum $\lambda_{\max}(x) = 2$ under maximal structural brittleness, while remaining strictly increasing and convex in $B(x)$. This yields a smooth convex map from $[0,1]$ to $[1,2]$ ensuring boundedness, monotonicity, and order preservation while amplifying high-brittleness cases. Since $\partial\lambda/\partial l > 0$ for all $l \in \mathcal{F}$, any increase in cluster brittleness inflates $\hat{u}(x)$ via (12). As the entire inflation procedure is label-free, conformal calibration and inference remain valid in the transformed $\hat{u}$-space, preserving finite-sample guarantees.

3.2 Prompt Acceptance and Response Set

Prompt-Level Decision Rule. For each prompt x, let define the returned response $\tilde{y} = y_{i^*}$ as the most representative member of dominant cluster $k^\star$, i.e., cluster with largest total assignment mass. We use the inflated uncertainty $\hat{u}(x) \in \mathbb{R}^{[0,1]}$ as a prompt-level score so that small $\hat{u}(x)$ indicates robust semantic agreement after inflation, while large $\hat{u}(x)$ indicates cluster brittleness. Since CP requires one label per prompt, we define a prompt-level error label by evaluating the returned response $E(x) = e(x, \tilde{y}) \in \{0,1\}$ where $E(x) = 0$ indicates that the returned response is correct. Using $\mathcal{I}_0 = \{\hat{u}(x) : x \in \mathcal{D}_{\text{cal}} \wedge E(x) = 0\}$ as calibration prompts whose returned response is correct, we calibrate a prompt acceptance threshold $\hat{\tau} = \text{Quantile}_{1-\alpha}\big(\mathcal{I}_0\big)$ which is $(1-\alpha)$-quantile of scores $\hat{u}(.)$ in the multiset $\mathcal{I}_0$ (if $\mathcal{I}_0 = \emptyset$, we abstain on all prompts). At inference, to answer a prompt x, we

$$\text{accept } x \iff \hat{u}(x) \leq \hat{\tau}\,; \quad \text{abstain otherwise.} \tag{16}$$

This calibration is appropriate since it aligns the prompt-level score $\hat{u}(x)$ with the prompt-level LLM outcome, i.e., correctness of the actually returned response $\tilde{y}$.

Theorem 1 (Conformal coverage for prompt acceptance)**.** *Assume $\tilde{y}$ and $\hat{u}(\cdot)$ are computed by the same procedure for calibration and test prompts under exchangeability of $\mathcal{D}_{\text{cal}}$ and test prompt x_{new}. With the $(1-\alpha)$-quantile $\hat{\tau}$ defined on $\hat{u}$-space of prompts with correct returned responses, we have*

$$\mathbb{P}(\hat{u}(x_{\text{new}}) \leq \hat{\tau} \mid E(x_{\text{new}}) = 0) \geq 1 - \alpha\,. \tag{17}$$

Response-Level Prediction Sets. To construct CP-consistent prediction sets over sampled responses, we define a response-level non-conformity score that combines prompt uncertainty with response atypicality. Using the clustering outputs, we first define a response conformity measure

$$\phi_i \ = \ s_{i\hat{k}} \cdot \mathcal{P}(C_{\hat{k}}) \ \in \mathbb{R}^{[0,1]} \quad \text{s.t.} \quad \hat{k} = \arg\max_k s_{ik}\,, \tag{18}$$

which is large when response y_i is both strongly assigned to its best-matching semantic cluster and that cluster is well supported across samples. We then define the response-level non-conformity score (smaller is better) as,

$$S(x, y_i) \ = \ \frac{1}{2}\Big(\hat{u}(x) + \big(1 - \phi_i\big)\Big) \ \in \ \mathbb{R}^{[0,1]}\,, \tag{19}$$

which increases with prompt-level uncertainty $\hat{u}(x)$ and with response atypicality. Thus, a response attains a low score only when it arises from a low-uncertainty prompt and is semantically representative within a well-supported cluster. For calibration prompt-response pairs, we compute $S(x_j, y_{ji})$ in (19) and form the multiset $\mathcal{S}_0 = \{S(x_j, y_{ji}) : \ e_{ji} = 0\}$ of scores corresponding to correct responses. The $(1-\alpha)$-quantile threshold is then defined as $\hat{q} = \text{Quantile}_{1-\alpha}(\mathcal{S}_0)$. For a test prompt x with sampled responses $\mathcal{Y}(x) = \{y_i\}_{i=1}^{n}$, the response-level prediction set is

$$\mathcal{C}_\alpha(x) \ = \ \big\{\, y_i \in \mathcal{Y}(x) : \ S(x, y_i) \leq \hat{q} \,\big\}\,, \tag{20}$$

where the size of $\mathcal{C}_\alpha(x)$ provides a direct measure of response-level uncertainty. If a single output is required, we return the most representative response $\tilde{y} \in \mathcal{C}_\alpha(x)$.

Theorem 2 (Conformal coverage for prediction response sets)**.** *Assume $S(x, y)$ is computed identically on calibration and test triples (x, y, e) under exchangeability with the same response sampling procedure. Considering $\mathcal{C}_\alpha$ defined in* (20)*, for a randomly drawn test triple $(x_{\text{new}}, y_{\text{new}}, e_{\text{new}})$,*

$$\mathbb{P}(y_{\text{new}} \in \mathcal{C}_\alpha(x_{\text{new}}) \mid e_{\text{new}} = 0) \ \geq \ 1 - \alpha\,. \tag{21}$$

The proofs of the theorems and computational overhead analysis of our approach are reported in Appendices A and B of the extended version, respectively.

4 Experimental Evaluations

We evaluate the performance of ACSE using discriminative and conformal analyses, benchmarking its ability to identify hallucinations (incorrect responses) reliably through adaptive uncertainty inflation against both uncalibrated and calibrated baselines. Appendix C of the extended version describes the experimental setup and evaluation metrics.

Method	TriviaQA					CoQA					NQ					TruthfulQA					MMLU				
	AUROC	FPR@95	FPR@90	AUPR	AUARC	AUROC	FPR@95	FPR@90	AUPR	AUARC	AUROC	FPR@95	FPR@90	AUPR	AUARC	AUROC	FPR@95	FPR@90	AUPR	AUARC	AUROC	FPR@95	FPR@90	AUPR	AUARC
TE	0.65	0.78	0.72	0.63	0.61	0.60	0.85	0.79	0.54	0.58	0.62	0.82	0.75	0.57	0.60	0.55	0.88	0.81	0.50	0.53	0.52	0.90	0.84	0.50	0.52
P(True)	0.70	0.72	0.65	0.68	0.65	0.66	0.75	0.68	0.63	0.62	0.67	0.74	0.66	0.64	0.64	0.60	0.79	0.71	0.55	0.59	0.58	0.85	0.78	0.54	0.55
EigV	0.71	0.65	0.58	0.69	0.69	0.73	0.68	0.61	0.67	0.66	0.74	0.66	0.59	0.70	0.69	0.70	0.72	0.64	0.66	0.64	0.68	0.75	0.68	0.64	0.62
SU	0.73	0.55	0.49	0.73	0.72	0.74	0.60	0.53	0.71	0.70	0.75	0.58	0.50	0.72	0.73	0.71	0.65	0.57	0.68	0.68	0.69	0.69	0.61	0.65	0.65
DDCRP-CP	0.77	0.52	0.46	0.76	0.74	0.77	0.55	0.48	0.74	0.72	0.78	0.54	0.46	0.75	0.75	0.74	0.61	0.53	0.71	0.70	0.71	0.66	0.58	0.68	0.67
CAP	0.80	0.48	0.41	0.79	0.76	0.79	0.52	0.44	0.77	0.75	0.80	0.50	0.42	0.78	0.77	0.76	0.58	0.50	0.73	0.72	0.73	0.62	0.54	0.70	0.69
ACSE (Ours)	**0.88**	**0.31**	**0.28**	**0.86**	**0.85**	**0.87**	**0.37**	**0.32**	**0.84**	**0.83**	**0.84**	**0.36**	**0.29**	**0.81**	**0.87**	**0.82**	**0.43**	**0.36**	**0.77**	**0.76**	**0.80**	**0.46**	**0.39**	**0.73**	**0.72**

Table 1: Hallucination detection performance across diverse benchmarks using Mistral-7b. Metrics evaluate discrimination (AUROC ↑, AUPR ↑), safety (FPR@95 ↓, FPR@90 ↓), and selective generation (AUARC ↑).

Method	Mistral-7B					LLaMA-2-7B					Falcon-7B					Qwen-7B				
	AUROC	FPR@95	FPR@90	AUPR	AUARC	AUROC	FPR@95	FPR@90	AUPR	AUARC	AUROC	FPR@95	FPR@90	AUPR	AUARC	AUROC	FPR@95	FPR@90	AUPR	AUARC
TE	0.65	0.78	0.72	0.63	0.61	0.59	0.85	0.78	0.56	0.55	0.55	0.89	0.82	0.53	0.52	0.64	0.80	0.72	0.61	0.62
P(True)	0.70	0.72	0.65	0.68	0.65	0.65	0.77	0.69	0.61	0.61	0.61	0.81	0.74	0.58	0.59	0.70	0.72	0.65	0.67	0.68
EigV	0.71	0.65	0.58	0.69	0.69	0.69	0.66	0.59	0.66	0.67	0.67	0.72	0.64	0.62	0.64	0.74	0.61	0.54	0.71	0.72
SU	0.73	0.55	0.49	0.73	0.72	0.75	0.56	0.49	0.72	0.72	0.72	0.60	0.53	0.70	0.68	0.79	0.50	0.43	0.77	0.78
DDCRP-CP	0.77	0.52	0.46	0.76	0.74	0.77	0.52	0.45	0.74	0.74	0.74	0.57	0.49	0.71	0.72	0.82	0.47	0.40	0.79	0.81
CAP	0.80	0.48	0.41	0.79	0.76	0.79	0.48	0.41	0.76	0.76	0.76	0.54	0.46	0.73	0.74	0.84	0.43	0.36	0.81	0.83
ACSE (Ours)	**0.89**	**0.28**	**0.24**	**0.88**	**0.87**	**0.86**	**0.35**	**0.29**	**0.83**	**0.84**	**0.84**	**0.38**	**0.32**	**0.81**	**0.82**	**0.91**	**0.25**	**0.19**	**0.89**	**0.90**

Table 2: Hallucination detection performance on TriviaQA across diverse architectures. Metrics evaluate discrimination (AUROC ↑, AUPR ↑), safety (FPR@95 ↓, FPR@90 ↓), and selective generation (AUARC ↑).

Implementation Details. For each prompt, we generate $n = 10$ responses via nucleus sampling ($\eta = 0.9$) with temperature 0.3 to balance diversity and coherence and embed them using the sentence encoder all-MiniLM-L6. We perform HAC clustering with average linkage and cosine distance threshold $\epsilon = 0.35$. We set λ to equally weigh all brittleness features as $\mathbf{w} = \langle \frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5} \rangle$ to avoid any prioritization, and set the overconfidence penalty threshold τ_{ref} to the 75th percentile of the calibration base-uncertainty distribution. Each dataset uses 2000 samples split into disjoint calibration set (60%) and test set (40%). For Size-Stratified Coverage Violation (SSCV) evaluation, we define prediction set size strata as $[1\text{-}2, 3\text{-}5, 6\text{-}7, 8\text{-}10]$. The source codes are available at *https://github.com/tailabTMU/ACSE*.

Datasets. We evaluate on five benchmarks spanning open-domain retrieval, conversational QA, hallucination detection, and multi-domain reasoning: *TriviaQA* [Joshi *et al.*, 2017], *CoQA* [Reddy *et al.*, 2019], *Natural Questions (NQ)* [Kwiatkowski *et al.*, 2019], *TruthfulQA* [Lin *et al.*, 2022], and a stratified subset of *MMLU* [Wang *et al.*, 2024].

LLM Models. We use *Mistral-7B-Instruct* [Jiang *et al.*, 2024] which combines Grouped-Query Attention (GQA) and Sliding Window Attention (SWA) for strong performance with efficient inference. To assess model-agnosticism, we additionally evaluate *Llama-2-7B-Chat* [Touvron *et al.*, 2023], *Falcon-7B-Instruct* (Multi-Query Attention; MQA), and *Qwen-7B-Chat* [Bai *et al.*, 2023].

Baselines. We compare ACSE against six uncertainty methods spanning lexical, semantic, geometric, and calibrated approaches: *Token Entropy (TE)* [Vaswani *et al.*, 2017], *P(True)* [Kadavath *et al.*, 2022], *Semantic Uncertainty (SU)* [Kuhn *et al.*, 2023], and *EigV* [Lin *et al.*, 2024]. We also include two CP-based calibrated frameworks: *DDCRP-CP* [Kaur *et al.*, 2024] and *Conformal Abstention Policy (CAP)* [Tayebati *et al.*, 2025]; we emphasize comparisons among DDCRP-CP, CAP, and ACSE since they share a CP calibration protocol, enabling standardized evaluation.

4.1 Results and Interpretations

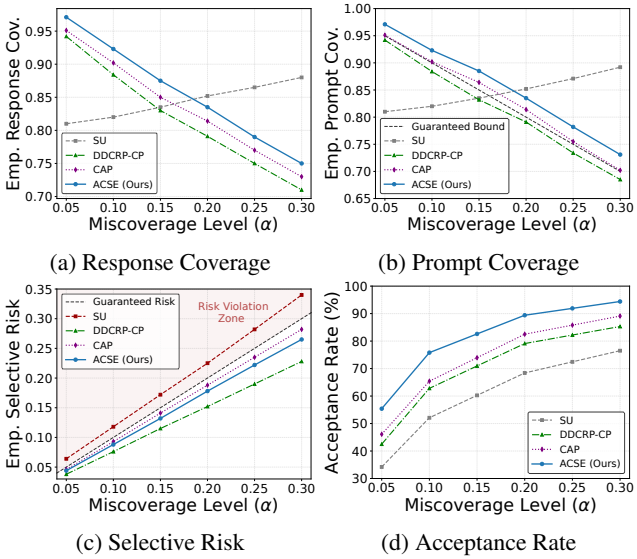
We conduct an extensive empirical evaluation of ACSE to assess whether it (1) improves hallucination detection relative to state-of-the-art baselines, (2) strictly enforces user-specified coverage via conformal mechanisms along with the selective risk while maintaining high acceptance rates, and (3) improves uncertainty quantification on hallucinations against baselines.

Hallucination Detection Performance. This evaluation assesses the ACSE's capability to distinguish between factually correct response generations and hallucinations across various datasets and LLMs. The results in Tables 1 and 2 demonstrate that ACSE achieves better discriminative performance across all benchmarks, e.g., on TriviaQA, ACSE increases AUROC to 0.88 from CAP's 0.80. This improvement is particularly notable on high-efficiency models such as Falcon-7B where ACSE reduces 0.48 FPR@95 observed by CAP to 0.31, a 35.4% relative decrease in accepted hallucinations. Low standard deviations (0.01 to 0.02) across five independent runs confirm the statistical significance of the results. We report ablation studies and additional results on a text summarization dataset, recent 8B LLMs and baselines, together with base SE comparisons, probabilistic calibration results, and statistical significance tests in Appendix D of the extended version.

Conformal Guarantees and Selective Risk. We assess ACSE reliability by varying α using both response-level and

Method	$\alpha = 0.05$ (Strict)						$\alpha = 0.10$ (Standard)						$\alpha = 0.20$ (Permissive)					
	Response-level			Prompt-level			Response-level			Prompt-level			Response-level			Prompt-level		
	R-Cov ↑	APS ↓	SSCV ↓	P-Cov ↑	Acc. ↑	Risk ↓	R-Cov ↑	APS ↓	SSCV ↓	P-Cov ↑	Acc. ↑	Risk ↓	R-Cov ↑	APS ↓	SSCV ↓	P-Cov ↑	Acc. ↑	Risk ↓
SU	0.823	**1.05**	0.081	0.810	34.2%	0.064	0.816	**1.02**	0.089	0.820	52.1%	0.118	0.808	**1.01**	0.115	0.852	68.4%	0.225
DDCRP-CP	0.941	1.95	0.046	0.942	42.5%	**0.038**	0.923	1.45	0.051	0.884	62.8%	**0.076**	0.810	1.22	0.074	0.791	79.1%	**0.152**
CAP	0.953	1.72	0.038	0.951	46.1%	0.046	0.902	1.32	0.042	0.917	65.4%	0.093	0.827	1.15	0.062	0.814	82.5%	0.188
ACSE (Ours)	**0.985**	1.32	**0.024**	**0.971**	**55.4%**	0.044	**0.939**	1.07	**0.030**	**0.923**	**75.8%**	0.088	**0.842**	1.08	**0.045**	**0.835**	**89.4%**	0.178

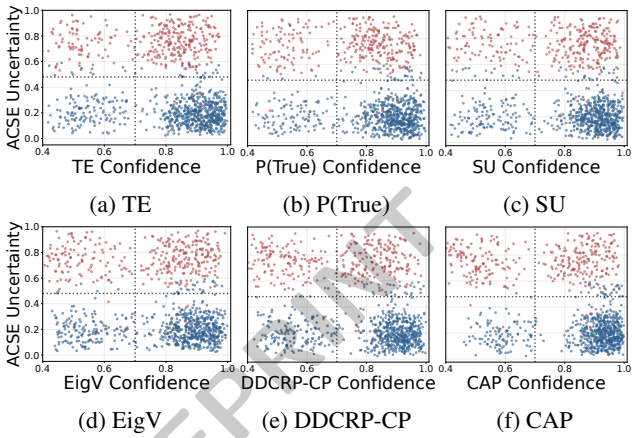
Table 3: Conformal analysis across varying α: Response Cov., APS, SSCV, Prompt Cov., Acceptance Rate (Acc.), and Selective Risk.



(a) Response Coverage (b) Prompt Coverage

(c) Selective Risk (d) Acceptance Rate

Figure 2: Sensitivity analysis against varying miscoverage level α.



(a) TE (b) P(True) (c) SU

(d) EigV (e) DDCRP-CP (f) CAP

Figure 3: Comparing ACSE uncertainty against baseline confidences, distinguishing correct responses (blue) from hallucinations (red).

prompt-level metrics reported in Table 3, to enforce user-specified guarantees through conformal decision rules against baselines. Statistical reliability is reported through Response-level Coverage (R-Cov) which measures the inclusion rate of correct generated responses in the prediction sets, and Prompt-level Coverage (P-Cov) which evaluates the fraction of accepted prompts with correct returned responses. The analyses in Figure 2a show that ACSE guarantees superior R-Cov across all α levels. Similarly, Figure 2b demonstrates that ACSE consistently satisfies the theoretical prompt coverage guarantee, whereas baselines frequently exhibit undercoverage violations. ACSE also maintains low Average Prediction Set Size (APS) of 1.07 compared to 1.32 APS of CAP at $\alpha = 0.10$. Additionally, ACSE achieves the lowest SSCV scores across all settings, e.g., 0.030 vs. 0.042 for CAP at $\alpha = 0.10$, indicating superior calibration stability across varying prediction set sizes. ACSE also consistently maximizes the Acceptance Rate (Acc.) across benchmarks, reaching 75.8% at $\alpha = 0.10$, a substantial improvement over the runner-up CAP at 65.4%. Furthermore, ACSE strictly guarantees the selective risk as the conditional error among accepted prompts that hallucinate, which is restricted by α as shown in Figure 2c. While the SU baseline suffers from overconfidence and risk violations, ACSE maintains a selective risk of 0.178 at low APS. ACSE also avoids the conservativeness of DDCRP-CP, which yields marginally lower risk but produces significantly larger prediction sets (1.45 vs. 1.07), trading precision for safety.

Uncertainty Quantification Analysis. This experiment evaluates the discriminative ability of our adjusted uncertainty score ($\hat{u}$) to separate correct responses from hallucinations. Figure 3 illustrates this separation across four quadrants against each baseline. The top-left quadrant denotes the ideal abstention region for hallucinations where they exhibit both low baseline confidence and high ACSE uncertainty; conversely, the bottom-right quadrant denotes the ideal acceptance region for the correct responses. ACSE consistently assigns high uncertainty to hallucinations and low uncertainty to correct responses whereas the baselines exhibit varying ranges of confidence while hallucinating. The top-right quadrant highlights baseline failures compared to ACSE, in which hallucinations are incorrectly assigned high confidence but are correctly flagged by ACSE with high uncertainty.

5 Conclusion

We introduced ACSE to mitigate LLM overconfidence by quantifying uncertainty through semantic dispersion. Rather than treating uncertainty as a one-dimensional signal, ACSE employs an adaptive inflation mechanism to penalize semantic brittleness, ensuring reliable estimates. Evaluations across diverse benchmarks demonstrate that ACSE significantly improves hallucination detection, discriminative performance, and coverage guarantees. ACSE offers a robust solution for the reliable deployment of LLMs in safety-critical domains. Future work will explore adaptive sampling strategies to reduce computational overhead and extend the framework to conformal risk control for fine-grained quality calibration.

Acknowledgments

This research was undertaken thanks in part to funding from the Canada First Research Excellence Fund at Toronto Metropolitan University and Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery grants (#348100).

References

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

[Bai *et al.*, 2023] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

[Brown *et al.*, 2020] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[Cer *et al.*, 2018] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*, 2018.

[Chang *et al.*, 2024] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3):1–45, 2024.

[Duan *et al.*, 2023] 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. *arXiv preprint arXiv:2307.01379*, 2023.

[Fadeeva *et al.*, 2024] Ekaterina Fadeeva, Aleksandr Rubashevskii, Artem Shelmanov, Sergey Petrakov, Haonan Li, Hamdy Mubarak, Evgenii Tsymbalov, Gleb Kuzmin, Alexander Panchenko, Timothy Baldwin, et al. Fact-checking the output of large language models via token-level uncertainty quantification. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9367–9385, 2024.

[Gao *et al.*, 2021] Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*, 2021.

[Holtzman *et al.*, 2020] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations (ICLR)*, 2020.

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

[Jiang *et al.*, 2024] AQ Jiang, A Sablayrolles, A Mensch, C Bamford, DS Chaplot, Ddl Casas, F Bressand, G Lengyel, G Lample, L Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2024.

[Joshi *et al.*, 2017] Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 1601–1611, 2017.

[Kadavath *et al.*, 2022] 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.

[Kaur *et al.*, 2024] Ramneet Kaur, Colin Samplawski, Adam D. Cobb, Anirban Roy, Brian Matejek, Manoj Acharya, Daniel Elenius, Alexander M. Berenbeim, John A. Pavlik, Nathaniel D. Bastian, and Susmit Jha. Addressing uncertainty in LLMs to enhance reliability in generative AI, 2024.

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

[Kwiatkowski *et al.*, 2019] 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.

[Lin *et al.*, 2022] Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th annual meeting of the association for computational linguistics*, volume 1, pages 3214–3252, 2022.

[Lin *et al.*, 2024] Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Generating with confidence: Uncertainty quantification for black-box large language models, 2024.

[Liu *et al.*, 2023] Yiheng Liu, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, Antong Li, Mengshen He, Zhengliang Liu, et al. Summary of chatgpt-related research and perspective towards the future of large language models. *Meta-radiology*, 1(2):100017, 2023.

[Murtagh and Contreras, 2017] Fionn Murtagh and Pedro Contreras. Algorithms for hierarchical clustering: an overview, ii. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 7(6):e1219, 2017.

[Reddy *et al.*, 2019] Siva Reddy, Danqi Chen, and Christopher D Manning. CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266, 2019.

[Reimers and Gurevych, 2019] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

[Shanahan, 2024] Murray Shanahan. Talking about large language models. *Communications of the ACM*, 67(2):68–79, 2024.

[Shorinwa *et al.*, 2025] Ola Shorinwa, Zhiting Mei, Justin Lidard, Allen Z Ren, and Anirudha Majumdar. A survey on uncertainty quantification of large language models: Taxonomy, open research challenges, and future directions. *ACM Computing Surveys*, 2025.

[Tayebati *et al.*, 2025] Sina Tayebati, Divake Kumar, Nastaran Darabi, Dinithi Jayasuriya, Ranganath Krishnan, and Amit Ranjan Trivedi. Learning conformal abstention policies for adaptive risk management in large language and vision-language models. *arXiv preprint arXiv:2502.06884*, 2025.

[Teubner *et al.*, 2023] Timm Teubner, Christoph M Flath, Christof Weinhardt, Wil Van Der Aalst, and Oliver Hinz. Welcome to the era of chatgpt et al. the prospects of large language models. *Business & Information Systems Engineering*, 65(2):95–101, 2023.

[Thirunavukarasu *et al.*, 2023] Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. Large language models in medicine. *Nature medicine*, 29(8):1930–1940, 2023.

[Touvron *et al.*, 2023] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[Vaswani *et al.*, 2017] 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.

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

[Wang *et al.*, 2024] Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290, 2024.

[Yadkori *et al.*, 2024] Yasin Abbasi Yadkori, Ilja Kuzborskij, András György, and Csaba Szepesvári. To believe or not to believe your LLM. *arXiv preprint arXiv:2406.02543*, 2024.

[Yao *et al.*, 2019] Jiayu Yao, Weiwei Pan, Soumya Ghosh, and Finale Doshi-Velez. Quality of uncertainty quantification for bayesian neural network inference. *arXiv preprint arXiv:1906.09686*, 2019.

[Yarie *et al.*, 2024] Liz Yarie, Dominic Soriano, Leonard Kaczmarek, Benjamin Wilkinson, and Eduardo Vasquez. Mitigating token-level uncertainty in retrieval-augmented large language models. *Authorea Preprints*, 2024.