

# A Survey of Uncertainty Estimation in LLMs: Theory Meets Practice

Hsiu-Yuan Huang<sup>1,2</sup>, Yutong Yang<sup>1,2</sup>, Zhaoxi Zhang<sup>1,3</sup>, Sanwoo Lee<sup>1,2</sup>, Yunfang Wu<sup>1,2\*</sup>

<sup>1</sup>National Key Laboratory for Multimedia Information Processing, Peking University

<sup>2</sup>School of Computer Science, Peking University, Beijing, China

<sup>3</sup>School of Computer Science & Technology, Beijing Institute of Technology, Beijing, China

{huang.hsiuyuan}@stu.pku.edu.cn, {yytpku, sanwoo, wuyf}@pku.edu.cn, {1120210536}@bit.edu.cn

## Abstract

As large language models (LLMs) continue to evolve, understanding and quantifying the uncertainty in their predictions is critical for enhancing application credibility. However, the existing literature relevant to LLM uncertainty estimation often relies on heuristic approaches, lacking systematic classification of the methods. In this survey, we clarify the definitions of uncertainty and confidence, highlighting their distinctions and implications for model predictions. On this basis, we integrate theoretical perspectives—including Bayesian inference, information theory, and ensemble strategies—to categorize various classes of uncertainty estimation methods derived from heuristic approaches. Additionally, we address challenges that arise when applying these methods to LLMs. We also explore techniques for incorporating uncertainty into diverse applications, including out-of-distribution detection, data annotation, and question clarification. Our review provides insights into uncertainty estimation from both definitional and theoretical angles, contributing to a comprehensive understanding of this critical aspect in LLMs. We aim to inspire the development of more reliable and effective uncertainty estimation approaches for LLMs in real-world scenarios.

Moreover, uncertainty estimation can play a critical role in mitigating hallucinations in LLMs by providing an indication when answering questions outside their knowledge boundary (Li et al., 2023; Huang et al., 2023; Xu et al., 2024). Without effective measures of uncertainty in transformer-based systems, relying on generated language as a trustworthy source of information becomes difficult (Kuhn et al., 2023).

Despite the mature theoretical frameworks and practical applications for uncertainty estimation established in machine learning, these frameworks are typically model-specific (Banerjee et al., 2024) or may not be adaptable to emerging LLMs due to two main factors. First, LLMs encompass an immense number of parameters, which makes the computational costs of traditional uncertainty estimation methods prohibitively high (Arteaga et al., 2024). Second, the widespread use of commercial black-box API models complicates matters further, as these models often lack transparency and provide no access to internal parameters or output probabilities (Xiong et al., 2024). As a result, traditional uncertainty estimation approaches become impractical (Lin et al., 2024). Therefore, there is an urgent need for a new, generalized, and reliable uncertainty estimation scheme tailored for LLMs.

Recently, there have been several review articles on uncertainty estimation methods (Gawlikowski et al., 2023; Geng et al., 2024). However, many of these papers explore uncertainty estimation in a heuristic manner, focusing on issues like hallucination (Zhang et al., 2023b), or may not clearly differentiate between confidence and uncertainty, which can lead to misunderstandings in the field.

We argue that grounding these methods in a clear theoretical framework is crucial for helping readers fully understand the concepts and inspiring future researchers to address the challenges mentioned above. Therefore, this review aims to bridge that gap by offering a more theory-driven exploration

## 1 Introduction

As large language models (LLMs) continue to proliferate across various applications, understanding and quantifying their uncertainty has become increasingly important. Uncertainty estimation provides valuable insights into the confidence of model predictions, which is crucial for decision-making in high-stakes fields such as medical diagnosis (Fox, 1980; Simpkin and Schwartzstein, 2016), where incorrect predictions can have serious consequences (Alkaissi and McFarlane, 2023; Shen et al., 2023).

\* Corresponding author.

of uncertainty estimation methods.

We begin by clarifying some easily confused concepts (Section 2). Next, we introduce the cornerstone of uncertainty estimation: Bayesian inference (Section 3). These methods rely on modeling the distributions of model parameters, which makes them not directly applicable to LLMs. However, it is still possible to indirectly incorporate Bayesian ideas for uncertainty estimation through various approximation techniques, often heuristic in nature. Following this, we discuss ensemble strategy (Section 4), a non-Bayesian approach commonly used to approximate distributions, and we tailor this concept specifically for LLMs. Then, we illustrate the uncertainty in LLMs through the lens of information theory (Section 5), using entropy, perplexity, and mutual information. Additionally, we explore existing verbal-based approaches to uncertainty estimation from the perspective of language expression (Section 6), a unique characteristic of LLMs. These multi-faceted perspectives allow us to present a comprehensive understanding of uncertainty estimation in LLMs, bridging the gap between theoretical foundations and practical methods. In Section 7, we demonstrate the task highly related to uncertainty. Figure 1 provides an overview of the article’s structure.

## 2 Preliminary

### 2.1 Uncertainty: Aleatoric vs. Epistemic

According to Kiureghian and Ditlevsen (2009), there are two main types of uncertainty: epistemic (systematic) uncertainty that is caused by exceeding knowledge boundaries or lack of data, which can be reduced by expanding the training data; aleatoric (statistical) uncertainty that captures the inherent randomness within the experiment, which is inevitable in nature.

Although there is no consensus on whether these two types of uncertainty (epistemic and aleatoric) should be strictly separated in machine learning (Hüllermeier and Waegeman, 2021), clarifying the distinction can help us better understand the challenges of uncertainty estimation in LLMs. (1) Since the training data for LLMs is either unknown or too vast to retrieve realistically, confirming whether a specific piece of knowledge falls outside the model’s learned scope is impractical, making the evaluation of epistemic uncertainty difficult. (2) Additionally, the varying decoding strategies used in LLM generation complicate the evaluation

of aleatoric uncertainty, as the noise inherent in the generation process is harder to quantify.

According to Lambert et al. (2024), many existing uncertainty estimation works (61.95%) assess total uncertainty rather than distinguishing between specific types. This paper primarily addresses total uncertainty, however, we also highlight the importance of differentiating between uncertainty types in specific contexts, such as question clarification (Section 7.3). It is worth noticing that in the realm of LLM, addressing and mitigating epistemic uncertainty should be our ultimate goal. A reliable AI assistant should be able to recognize when a situation exceeds its knowledge boundaries and either prompt human intervention or refuse to provide answers.

### 2.2 Differences: Uncertainty vs. Confidence

Uncertainty and confidence are distinct yet inter-related aspects of model evaluation, particularly in LLMs. While some researchers suggest that increased uncertainty correlates with decreased confidence (Geng et al., 2024; Xiao et al., 2022; Chen and Mueller, 2023), this view lacks a clear distinction between the two concepts. Following Lin et al. (2024), for  $P(Y|x) = \mathcal{N}(\mu, \sigma^2)$ ,  $\sigma^2$  represents uncertainty, while confidence in output  $Y = y_0$  is expressed as  $-\frac{y_0 - \mu}{\sigma}$ . For instance, in a classification task, low uncertainty signifies a dominant class probability, which correlates with high confidence. However, high confidence does not necessarily imply low uncertainty, as the probabilities of other classes contribute to the overall uncertainty.

In brief, uncertainty refers to the overall output distribution, indicating the variability in potential predictions, whereas confidence pertains to a specific prediction, denoting the likelihood of that output (Manakul et al., 2023). The existing works mainly focus on aggregating **multiple responses** to get the most accurate answer, as shown in Section 4, than focus on one particular prediction. Therefore, we argue that “uncertainty” is a more precise term than “confidence” in the context of estimation. In this article, we emphasize uncertainty estimation.

## 3 Uncertainty Estimation with Bayesian Inference

The Bayesian theory is crucial for estimating uncertainty, as most methodologies draw upon concepts derived from Bayesian theory to varying degrees.

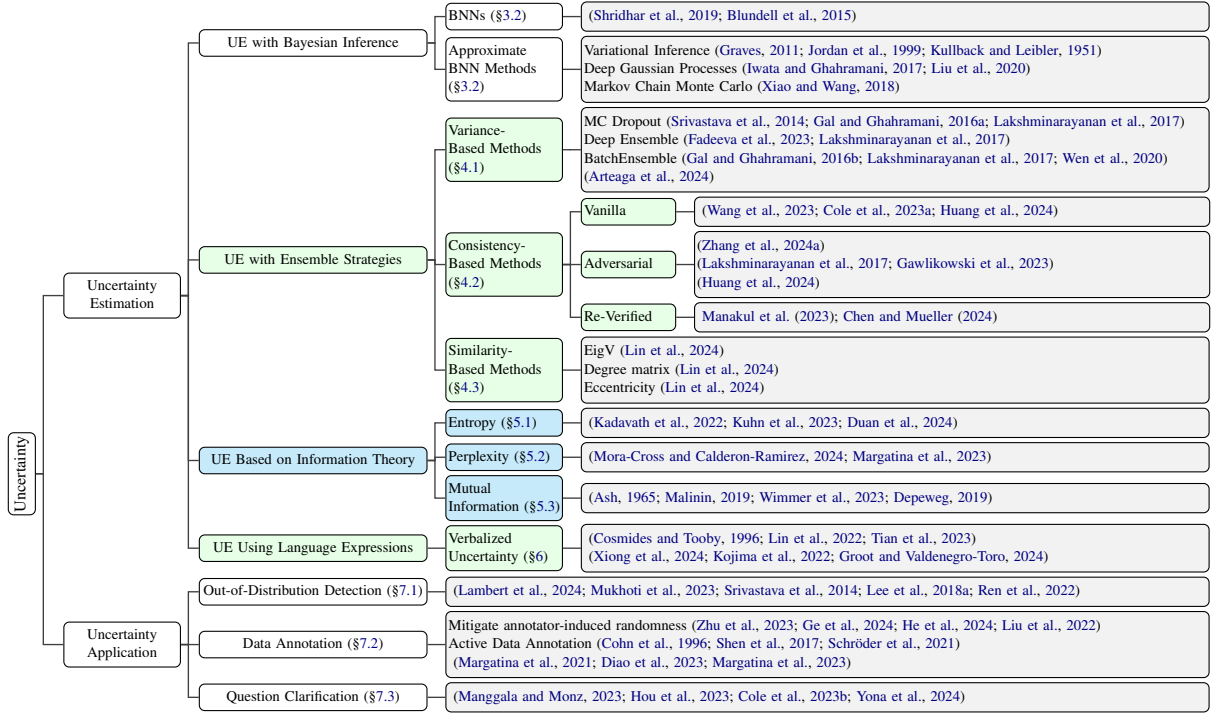


Figure 1: Taxonomy of uncertainty estimates. Blue nodes stand for white-box-LLMs-only methods, while green nodes stand for methods suitable for black-box LLMs as well.

### 3.1 Bayesian Neural Networks

Bayesian Neural Networks (BNNs) estimate the posterior distribution of weights  $p(w|\mathcal{D})$  based on training data  $\mathcal{D}$ , which is crucial in Bayesian inference as it represents the updating of beliefs about model weights with increasing data (Shridhar et al., 2019). By taking the expectation over the posterior, BNNs calculate the predictive distribution of label  $y$  given input  $x$  as  $\mathbb{E}_{p(w|\mathcal{D})} [P(y|x, w)]$ , effectively incorporating uncertainty into their predictions. In contrast to traditional Deep Neural Networks (DNNs) that rely on point estimates, typically optimized via maximum likelihood or maximum a posteriori estimation, BNNs provide a more comprehensive view of model uncertainty.

### 3.2 Approximate BNN Methods

While BNNs provide a principled framework for uncertainty estimation, the complexity and computational cost of obtaining the exact posterior is often unacceptable. To address this, approximate BNN methods have been developed to efficiently approximate the posterior distribution.

**Variational Inference** Variational Inference (VI) is a crucial approach for estimating model uncertainty (Graves, 2011), which involves constructing a simpler variational distribution to approximate the true posterior (Jordan et al., 1999). Theoretically, the Kullback-Leibler (KL) divergence (Kull-

back and Leibler, 1951) measures the difference between these two distributions. However, computing the KL divergence directly is often intractable. Instead, the Evidence Lower Bound (ELBO) (Jordan et al., 1999) is more commonly employed to approximate the posterior.

**Deep Gaussian Processes** Deep Gaussian Processes (DGP) are flexible, non-parametric models that leverage Bayesian inference to predict outcomes by estimating posterior distributions from prior data (Iwata and Ghahramani, 2017). Liu et al. (2020) introduces the concept of distance awareness, which measures the similarity between inference samples and training data. While DGPs are adaptable and perform well with small datasets, they come with high computational costs and scalability challenges. Moreover, the choice of kernel function is critical and requires careful tuning to optimize performance. DGPs are adaptable and perform well with small datasets but are computationally expensive and challenging to scale. The choice of kernel function is critical and requires tuning experience.

**Markov Chain Monte Carlo** Markov Chain Monte Carlo (MCMC) generates samples that approximate posterior distribution by constructing a sequence of states in a Markov chain. Each new state relies solely on the current state, ensuring that the process retains the Markov property. As the

chain progresses, the samples converge to the desired posterior distribution. [Xiao and Wang \(2018\)](#) applies MCMC in the context of semantic segmentation, demonstrating its utility.

### 3.3 Takeaways

BNNs’ inference provides both mean and variance, which are helpful in tasks needing high-confidence predictions like autonomous driving and medical diagnosis. Although BNNs offer a flexible framework for defining priors, selecting an appropriate prior for specific tasks remains challenging. Noise Contrastive Priors (NCP) ([Hafner et al., 2019](#)) proposes a method by adding noise to input data using a prior, and a wide distribution as the output prior. Considering the resource requirements, variational inference is the most ubiquitous method.

## 4 Uncertainty Estimation with Ensemble Strategies

Ensemble methods offer a robust framework to enhance predictive performance and, more pertinent to our focus, to provide uncertainty estimations. In this section, we summarize three mainstream approaches for integrating model outputs to assess uncertainty, including variance-based, consistency-based and similarity-based ensemble techniques.

### 4.1 Variance-Based Ensemble

Ensemble-based methods for DNNs measure uncertainty by leveraging the diversity of predictions generated by multiple model versions under slightly varied conditions ([Fadeeva et al., 2023](#)). Specifically, the *variance* between the predictions of the individual predictor can be seen as a natural vehicle for uncertainty estimation, with higher disagreement indicating higher uncertainty. The final prediction is calculated by averaging each model’s output ([Lakshminarayanan et al., 2017](#)), using:

$$p(y|\mathbf{x}) = M^{-1} \sum_{m=1}^M p_{\theta_m}(y|\mathbf{x}, \theta_m) \quad (1)$$

where  $M$  indicates the set of models, and the variance can be calculated as :

$$\sigma_*^2(\mathbf{x}) = M^{-1} \sum_m (\sigma_{\theta_m}^2(\mathbf{x}) + \mu_{\theta_m}^2(\mathbf{x})) - \mu_*^2(\mathbf{x}) \quad (2)$$

**Monte Carlo Dropout** Dropout ([Srivastava et al., 2014](#)) is a widely used regularization technique in neural networks, designed to prevent overfitting during the training phase by randomly discarding a

fraction of neurons. Monte Carlo Dropout (MCD) ([Gal and Ghahramani, 2016a](#)) extends this concept to the inference phase. By retaining the dropout mechanism during inference and randomly dropping neurons at each iteration, MCD generates multiple output samples. These samples can then be used to estimate the variance of the predictive distribution, providing valuable insights into the model’s uncertainty regarding its predictions.

**Deep Ensemble** Deep Ensemble, introduced by [Lakshminarayanan et al. \(2017\)](#), has emerged as a state-of-the-art method for uncertainty estimation. Unlike traditional single-model approaches, Deep Ensemble leverages multiple models trained independently on either randomly sampled subsets or the entire dataset ([Lakshminarayanan et al., 2017](#)).

**Batch Ensemble** Batch Ensemble ([Wen et al., 2020](#)), a more lightweight and parallelizable variant of deep ensemble, introduces the concepts of “slow weights” and “fast weights.” The slow weights serve as the base weights, while the fast weights provide modifications to these base weights, allowing Batch Ensemble to be effectively employed with LLMs ([Arteaga et al., 2024](#)).

### 4.2 Consistency-Based Ensemble

Research has shown that higher predictive uncertainty is associated with an increased likelihood of hallucinations ([Xiao and Wang, 2021](#)). Uncertainty estimation methods for black-box LLMs largely rely on heuristic approaches to mitigate hallucinations, linking uncertainty to both *confidence* and *hallucination scores* ([Zhang et al., 2023a](#); [Manakul et al., 2023](#)). We propose framing these heuristic methods within uncertainty estimation approaches from an ensemble perspective by focusing on how to evaluate response consistency, without strictly distinguishing between these concepts.

**Vanilla Methods** Inspired by [Chen and Mueller \(2024\)](#), we categorize methods for deriving uncertainty estimates from **user observations** of LLM responses as vanilla methods. The idea that answer consistency reflects uncertainty in LLMs is supported by [Wang et al. \(2023\)](#), who propose self-consistency (or temperature sampling) and find that consistency correlates highly with accuracy. This suggests that low consistency indicates uncertain and therefore confers some ability for the model to “know when it doesn’t know”. [Cole et al. \(2023a\)](#) introduced two metrics—*repetition* and *di-*



versity—to measure consistency. Let  $O$  denote the outputs of the LLM,  $O_g$  represent the greedy output, and  $S$  be the set of sampled outputs. Repetition is expressed as:

$$\text{Repetition} = \frac{|\{o \in S \mid o = O_g\}|}{|S|} \quad (3)$$

Diversity, inversely proportional to the number of distinct samples, is defined by:

$$\text{Diversity} = 1 - \frac{|S|}{|\text{Distinct}(S)|} \quad (4)$$

Here,  $|\text{Distinct}(S)|$  is the count of unique samples in  $S$ ; a value of zero is assigned if all samples are distinct. Additionally, Huang et al. (2024) framed uncertainty quantification in LLMs as a binary problem, where the LLM inconsistency is considered to be *uncertain* and vice versa *certain* after a definite number of samplings.

**Adversarial Methods** Maximizing diversity among individual networks is crucial when applying ensemble methods (Lakshminarayanan et al., 2017; Renda et al., 2019; Gawlikowski et al., 2023). The adversarial methods we summarised build upon vanilla approaches by adding a twist of adversarial components, which increase variability in LLM responses. Zhang et al. (2024a) introduce a mechanism that perturbs semantically equivalent questions to evaluate the consistency of LLM responses across variations of the same question. Additionally, Huang et al. (2024) inject correct and incorrect labels, respectively, into the prompt during sampling, in addition to using the vanilla method. The uncertainty level is then determined by the LLM responses’ consistency across three samplings for each instance. If the results are consistent, the model is classified as *certain*; otherwise *uncertain*.

**Re-Verified Methods** The biggest difference between re-verified and other consistency-based methods lies in the round of interactions. These methods check the LLM’s consistency by having the LLM itself, or another model, answer a closed-ended fact-checking question. While they are typically designed to mitigate LLM hallucinations, they can also be regarded as a means of evaluating consistency. Manakul et al. (2023); Chen and Mueller (2024) instructed the LLM to evaluate its response by selecting from a limited set of options, such as A) Correct, B) Incorrect, or C) Unsure, and then a numerical score is assigned to each option, and the average score across multiple rounds of such

verification questions is computed to determine the LLM’s uncertainty for each instance.

### 4.3 Similarity-Based Ensemble

These methods, proposed by Lin et al. (2024), calculate the similarity between multiple responses to indirectly quantify the dispersion of model outputs. Compared to consistency-based methods, similarity-based approaches offer a more continuous measurement of uncertainty, as the similarity is derived from the predicted probabilities of an off-the-shelf Natural Language Inference (NLI) model rather than by assigning values to different variables and averaging them. Lin et al. (2024) used a small NLI model to divide the semantic set of responses, where the predicted probabilities are viewed as the similarity. With the similarity between each answer obtained by the NLI model, we can construct the adjacency matrix. Then, the sum of eigenvalues of the graph Laplacian (EigV), degree matrix, and eccentricity can be calculated. For more detail, please refer to Lin et al. (2024).

### 4.4 Takeaways

Monte Carlo Dropout (Gal and Ghahramani, 2016b) and Deep Ensembles (Lakshminarayanan et al., 2017) are two classical white-box variance-based ensemble techniques that efficiently approximate Bayesian inference for uncertainty estimation. However, modifying multiple models can become prohibitively expensive (Lakshminarayanan et al., 2017), especially as the number of parameters scales up with LLMs, which renders these techniques primarily applicable in theory rather than practice for LLMs. In contrast, Batch Ensemble provides a practical solution to address this challenge.

As for the consistency or similarity based ensemble methods, they estimate uncertainty by evaluating the consistency or similarity of responses across multiple samples. While relatively straightforward, these methods come in various forms, each offering unique perspectives on uncertainty.

## 5 Uncertainty Estimation Based on Information Theory

Shannon established the foundations of information theory by integrating the principles of probability theory with the quantification of information, significantly impacting various fields, including machine learning. This section explores uncertainty

estimation in LLMs from the information theory point of view.

### 5.1 Entropy

For classification tasks, entropy indicates the degree of dispersion in the distribution of the model’s predictions for a given input (Wang et al., 2022; Malinin and Gales, 2018), which can be present in:

$$H(X) = - \sum_{i=1}^n p(x_i) \log p(x_i) \quad (5)$$

where  $p_i$  is the model’s prediction probability for category  $i$ . The higher the entropy value, the more uncertain the model’s prediction. Quantifying uncertainty across an entire sequence in text generation tasks using LLMs is more complex. Kadavath et al. (2022) use the probability of the complete sequence to compute Predictive Entropy (PE). Given an input  $x$  and a generated sentence  $s$  consisting of  $N$  tokens, where  $s$  is a completion based on  $x$ , the probability of generating the  $i$ -th token  $s_i$  given the preceding tokens  $s_{<i}$  and the prompt  $x$  is denoted as  $p(z_i \mid s_{<i}, x)$ . The **PE** of the entire sentence  $s$  is given by:

$$\mathbf{PE}(s, x) = - \log p(s|x) = \sum_i - \log p(s_i | s_{<i}, x) \quad (6)$$

This token-level measure is commonly used as a baseline for assessing uncertainty.

However, free-form text generation presents unique challenges due to semantic equivalence, where different sentences can convey the same meaning. This can lead to inflated uncertainty estimates at the token level because different tokens might represent similar meanings. To address this, **Semantic Entropy (SE)** has been proposed (Kuhn et al., 2023). SE enhances token-level measures by clustering sentences into equivalence classes based on their semantic similarity and computing entropy over these classes:

$$SE(x) = - \sum_c p(c|x) \log p(c|x) \quad (7)$$

where  $c$  denotes an equivalence class of semantically similar sentences.

Recent work by Duan et al. (2024) highlights that not all tokens contribute equally to the underlying meaning, as linguistic redundancy often allows a few key tokens to capture the essence of longer sentences. To improve uncertainty estimation, Duan et al. (2024) introduce **SAR** (Shifting Attention to Relevance), a heuristic method that

adjusts attention to more relevant components at both the token and sequence levels.

### 5.2 Perplexity

Perplexity is a widely used metric for evaluating the readability and accuracy of text generated by LLMs, which is calculated by exponentiating the average cross-entropy loss. It can be understood as a measure of how surprised the model is when evaluating a sequence of tokens (Mora-Cross and Calderon-Ramirez, 2024). A higher perplexity value indicates that the model assigns a more dispersed probability distribution, signifying greater uncertainty about the next word. Therefore, perplexity can also serve as an indicator of LLM uncertainty. Margatina et al. (2023) uses perplexity to evaluate the uncertainty of each in-context example and select those with the highest perplexity as few-shot data for in-context learning (ICL).

### 5.3 Mutual Information

A basic result from information theory shows that Shannon entropy can be additively decomposed into conditional entropy and mutual information (Ash, 1965):

$$H(Y) = H(Y|\Theta) + I(Y, \Theta) \quad (8)$$

Conditional entropy  $H(Y|\Theta)$  represents the uncertainty remaining in  $Y$  when the realization of  $\Theta$  is known, and it naturally serves as a measure of aleatoric uncertainty. Thus, mutual information, which captures the difference between total uncertainty and aleatoric uncertainty, serves as a measure of epistemic uncertainty (Depeweg, 2019).

Malinin (2019) posits that mutual information measures the ‘disagreement’ between models in an ensemble, and therefore reflects epistemic uncertainty, which arises from the model’s lack of understanding of the data. However, Wimmer et al. (2023) argue that mutual information is better interpreted as a measure of divergence or conflict rather than ignorance, may not be the right measure of epistemic uncertainty.

### 5.4 Takeaways

Notably, information-based methods require access to token-level probabilities from LLMs, which makes them unsuitable for current black-box LLMs or API models. A potential solution is to utilize white-box LLMs as surrogate models to provide these probabilities. Shrivastava et al. (2023) demonstrates that probabilities from weaker white-box

surrogate models can effectively estimate the internal confidence levels of stronger black-box models, such as GPT-4, and outperform linguistic uncertainty measures.

## 6 Uncertainty Estimation Using Language Expressions

By training on vast amounts of natural language data, current LLMs are able to produce language that closely approximates human speech. A key aspect of human intelligence lies in our capability to express and communicate our uncertainty in a variety of ways (Cosmides and Tooby, 1996). Unlike uncertainty in the statistical sense, which relates to the degree of dispersion in output, this section examines LLM’s uncertainty through the lens of naturalistic expressions.

This scope of research focuses on prompting LLMs to explicitly articulate their level of uncertainty alongside their responses. Lin et al. (2022) introduces the concept of verbalized confidence that prompts LLMs to express its uncertainty using natural language for representing degrees. Tian et al. (2023) proposes prompting LLMs to generate the top-k guesses and their corresponding confidence for a given question. Xiong et al. (2024) proposes Self-Probing which is to ask LLMs “How likely is the above answer to be correct?” and have them verbalize its uncertainty in the form of the numerical number in the range of 0-100%.

It is important to note that the conclusions regarding uncertainty estimation in this section vary by methods, tasks, evaluation metrics, and models, and these findings should be interpreted within their specific context. Xiong et al. (2024) conducts an empirical assessment of zero-shot verbal-based methods across different sampling and aggregation strategies, revealing that LLMs often exhibit overconfidence. This overconfidence can be mitigated by employing prompting strategies, such as zero-shot Chain of Thought (CoT) (Kojima et al., 2022). However, Tian et al. (2023) reaches an opposite conclusion, indicating that CoT does not enhance verbalized calibration. Additionally, research has shown that the accuracy of verbal-based uncertainty estimation varies by task. For instance, in sentiment analysis, models tend to be underconfident, while overconfidence is observed in tasks such as math word problems and named entity recognition (Groot and Valdenegro-Toro, 2024).

## 7 Uncertainty Application

### 7.1 Out-of-Distribution Detection

Out-of-distribution (OOD) data refers to samples that significantly deviate from the training data of a machine learning model (Lambert et al., 2024). In other words, when the model encounters inputs beyond its training knowledge, it typically exhibits high uncertainty. Epistemic uncertainty, which reflects the model’s lack of knowledge about certain inputs, is interrelated with and complementary to OOD detection. On the one hand, by leveraging epistemic uncertainty, the accuracy of OOD detection can be enhanced, as high epistemic uncertainty often indicates that the sample is likely to be OOD (Mukhoti et al., 2023). On the other hand, understanding OOD data can help refine uncertainty estimation methods. By recognizing the high uncertainty models exhibited when faced with OOD data, researchers have developed techniques like Monte Carlo Dropout (Srivastava et al., 2014), significantly improving OOD detection.

Mahalanobis Distance (MD) is widely employed for OOD detection. Given a mean vector  $\mu$  and a covariance matrix  $\Sigma$ , the MD is defined as:

$$MD(z_{test}; \Sigma, \mu) = (z_{test} - \mu)^T \Sigma^{-1} (z_{test} - \mu) \quad (9)$$

The method (Lee et al., 2018a) calculates uncertainty using MD based on the distance to the nearest class-conditional Gaussian distribution. Greater deviation indicates higher uncertainty, suggesting the sample is likely OOD. Notably, MD can also be considered a density-based uncertainty estimation method (Fadeeva et al., 2023). Both OOD detection and density-based uncertainty measurement aim to identify data points that differ from the training data. In OOD detection, these points are treated as outliers, whereas in density-based approaches, they are often seen as residing in low-density regions, which correlate with higher uncertainty.

Based on MD, Ren et al. (2022) introduced the Relative Mahalanobis Distance (RMD), defined as:

$$MD(z_{test}) := MD(z_{test}; \mu^z, \Sigma^z) \quad (10)$$

$$RMD(z_{test}) := MD(z_{test}) - MD_0(z_{test}) \quad (11)$$

where  $MD_0$  represents the distance from a sample to the global distribution (Ren et al., 2022). MD is becoming popular as a new concept and has extended with multiple variants (Vazhentsev et al., 2023; Ren et al., 2023; Lee et al., 2018b).

## 7.2 Data Annotation

**Mitigate the Randomness Introduced by Annotators** Supervised learning fundamentally depends on manually labeled data, often referred to as “gold standard” annotations. However, human annotators are inherently susceptible to variability and subjective interpretation (Zhu et al., 2023), contributing to the aleatoric uncertainty.

Zhu et al. (2023) proposed a method using BNNs to detect annotation errors. In recognition of human cognitive bias, the authors introduced a non-zero value to represent the variance associated with these errors. Similarly, Ge et al. (2024) applied Monte Carlo Dropout (MCD) to balance annotation errors during the distillation process across multiple languages. Beyond error detection, uncertainty plays a critical role in other aspects of Named Entity Recognition (NER). Considering that NER is extremely sensitive to “gold standard” data He et al. (2024), recognizing and utilizing the data uncertainty plays a critical role in boosting its performance He et al. (2024); Liu et al. (2022).

Furthermore, uncertainty can be integrated with LLMs. Zhang et al. (2024b) introduced a method linking NER with ICL. In their approach, several smaller models first perform traditional NER. Then, the result with the lowest uncertainty is selected and incorporated into the prompt for ICL.

**Active Data Annotation** Active Learning (AL) (Cohn et al., 1996) aims to enhance data labeling efficiency by identifying the most informative unlabeled data for annotation within reasonable budget constraints. Uncertainty can serve as a key metric for determining which data points are most valuable to annotate.

For traditional supervised active learning, uncertainty sampling is widely considered one of the most effective approaches (Shen et al., 2017; Schröder et al., 2021; Margatina et al., 2021). However, in ICL scenarios, researchers have reported mixed results. For instance, Diao et al. (2023) demonstrated that selecting the most uncertain questions for annotation yielded promising results on eight widely used reasoning task datasets. In contrast, Margatina et al. (2023) found that uncertainty-based methods underperformed compared to other prevalent AL algorithms.

A potential explanation for this discrepancy is that larger models benefit from demonstrations with higher uncertainty, whereas smaller models, such as GPT-2 and GPT-large, perform better when

provided with low-uncertainty prompts.

## 7.3 Question Clarification

In real-world scenarios, queries often contain some degree of ambiguity due to missing background knowledge, insufficient context, or open-endedness. Recently, several works have employed various uncertainty-based methods to identify ambiguous questions and guide appropriate follow-up actions.

Manggala and Monz (2023) investigate the alignment between predictive uncertainty and ambiguous instructions in visually grounded communication. Specifically, they generated pairs of clear and ambiguous instructions through minimal edits (such as removing color or quantity information). Suppose the predictive uncertainty for an instruction is significantly higher than its clear counterpart. In that case, the instruction can be deemed ambiguous, at which point the model should ask clarifying questions to resolve the ambiguity.

Hou et al. (2023) proposes an input clarification ensembling framework that generates an ensemble of input clarifications. By evaluating the disagreement among these clarifications, the authors quantify aleatoric uncertainty (due to input ambiguity) arising from total ambiguity. Similarly, Cole et al. (2023b) measures denotational uncertainty—defined as question ambiguity—by generating various possible interpretations of a question and selecting the most likely interpretation.

On a different note, Yona et al. (2024) proposes the GRANOLA QA evaluation method, which allows multiple levels of granularity for correct answers. They introduced a new decoding strategy, Decoding with Response Aggregation (DRAG), which adjusts the granularity of answers based on the model’s level of uncertainty. If the model is uncertain about a specific answer, it may provide a coarser-grained response, which, while less informative, has a higher likelihood of being correct.

## 8 Conclusion

In this paper, we provide a comprehensive survey of uncertainty and its estimation from four distinct perspectives, offering valuable insights into uncertainty estimation in LLMs. Our work aims to bridge the gap between theoretical foundations and practical methodologies, contributing to a deeper understanding of uncertainty estimation in LLMs. Additionally, we showcase various methods for integrating uncertainty into a range of applications.



We hope to inspire the development of innovative approaches that enhance the reliability of LLMs across diverse contexts.

## Limitations

This review, while comprehensive, has certain limitations. First, the focus on four theoretical perspectives—Bayesian inference, information theory, ensemble strategies, and language expression—may overlook other emerging approaches that could contribute to uncertainty estimation. Furthermore, the rapidly evolving nature of the field means that new methodologies may emerge that are not covered in this review. Lastly, our survey and classification of methods is inherently subjective, as it is influenced by the selected literature and its interpretation.

## Ethics Statement

**Use of AI Assistants** We have employed ChatGPT as a writing assistant, primarily for polishing the text after the initial composition.

## References

- Hussam Alkaissi and Samy I McFarlane. 2023. Artificial hallucinations in chatgpt: implications in scientific writing. *Cureus*, 15(2).
- Gabriel Y. Arteaga, Thomas B. Schön, and Nicolas Pielawski. 2024. [Hallucination detection in llms: Fast and memory-efficient finetuned models](#). *Preprint*, arXiv:2409.02976.
- Robert B Ash. 1965. *Information theory*. Dover Publications.
- Sourav Banerjee, Ayushi Agarwal, and Saloni Singla. 2024. [Llms will always hallucinate, and we need to live with this](#). *Preprint*, arXiv:2409.05746.
- Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. 2015. Weight uncertainty in neural network. In *International conference on machine learning*, pages 1613–1622. PMLR.
- Jiuhai Chen and Jonas Mueller. 2023. [Quantifying uncertainty in answers from any language model and enhancing their trustworthiness](#). *Preprint*, arXiv:2308.16175.
- Jiuhai Chen and Jonas Mueller. 2024. [Quantifying uncertainty in answers from any language model and enhancing their trustworthiness](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5186–5200, Bangkok, Thailand. Association for Computational Linguistics.
- David A Cohn, Zoubin Ghahramani, and Michael I Jordan. 1996. Active learning with statistical models. *Journal of artificial intelligence research*, 4:129–145.
- Jeremy R. Cole, Michael J. Q. Zhang, Daniel Gillick, Julian Martin Eisenschlos, Bhuwan Dhingra, and Jacob Eisenstein. 2023a. [Selectively answering ambiguous questions](#). *Preprint*, arXiv:2305.14613.
- Jeremy R Cole, Michael JQ Zhang, Daniel Gillick, Julian Martin Eisenschlos, Bhuwan Dhingra, and Jacob Eisenstein. 2023b. Selectively answering ambiguous questions. *arXiv preprint arXiv:2305.14613*.
- Leda Cosmides and John Tooby. 1996. Are humans good intuitive statisticians after all? rethinking some conclusions from the literature on judgment under uncertainty. *cognition*, 58(1):1–73.
- Stefan Depeweg. 2019. [Modeling epistemic and aleatoric uncertainty with bayesian neural networks and latent variables](#).
- Shizhe Diao, Pengcheng Wang, Yong Lin, and Tong Zhang. 2023. Active prompting with chain-of-thought for large language models. *arXiv preprint arXiv:2302.12246*.
- Jinhao Duan, Renming Zhang, James Diffenderfer, Bhavya Kailkhura, Lichao Sun, Elias Stengel-Eskin, Mohit Bansal, Tianlong Chen, and Kaidi Xu. 2024. Gtbench: Uncovering the strategic reasoning limitations of llms via game-theoretic evaluations. *arXiv preprint arXiv:2402.12348*.
- Ekaterina Fadeeva, Roman Vashurin, Akim Tsvigun, Artem Vazhentsev, Sergey Petrakov, Kirill Fedyanin, Daniil Vasilev, Elizaveta Goncharova, Alexander Panchenko, Maxim Panov, et al. 2023. Lmpolygraph: Uncertainty estimation for language models. *arXiv preprint arXiv:2311.07383*.
- Renee C Fox. 1980. The evolution of medical uncertainty. *The Milbank Memorial Fund Quarterly. Health and Society*, pages 1–49.
- Yarin Gal and Zoubin Ghahramani. 2016a. [Dropout as a bayesian approximation: Representing model uncertainty in deep learning](#). In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1050–1059, New York, New York, USA. PMLR.
- Yarin Gal and Zoubin Ghahramani. 2016b. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR.
- Jakob Gawlikowski, Cedrique Rovile Njiteucheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, et al. 2023. A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, 56(Suppl 1):1513–1589.

- Ling Ge, Chunming Hu, Guanghui Ma, Jihong Liu, and Hong Zhang. 2024. [Discrepancy and uncertainty aware denoising knowledge distillation for zero-shot cross-lingual named entity recognition](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):18056–18064.
- Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koepl, Preslav Nakov, and Iryna Gurevych. 2024. [A survey of confidence estimation and calibration in large language models](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6577–6595, Mexico City, Mexico. Association for Computational Linguistics.
- Alex Graves. 2011. Practical variational inference for neural networks. *Advances in neural information processing systems*, 24.
- Tobias Groot and Matias Valdenegro-Toro. 2024. [Over-confidence is key: Verbalized uncertainty evaluation in large language and vision-language models](#). *Preprint*, arXiv:2405.02917.
- Danijar Hafner, Dustin Tran, Timothy Lillicrap, Alex Irpan, and James Davidson. 2019. [Noise contrastive priors for functional uncertainty](#). *Preprint*, arXiv:1807.09289.
- Jianfeng He, Linlin Yu, Shuo Lei, Chang-Tien Lu, and Feng Chen. 2024. [Uncertainty estimation on sequential labeling via uncertainty transmission](#). *Preprint*, arXiv:2311.08726.
- Bairu Hou, Yujian Liu, Kaizhi Qian, Jacob Andreas, Shiyu Chang, and Yang Zhang. 2023. Decomposing uncertainty for large language models through input clarification ensembling. *arXiv preprint arXiv:2311.08718*.
- Hsiu-Yuan Huang, Zichen Wu, Yutong Yang, Junzhao Zhang, and Yunfang Wu. 2024. [Unc-ttp: A method for classifying llm uncertainty to improve in-context example selection](#). *Preprint*, arXiv:2408.09172.
- Xiaowei Huang, Wenjie Ruan, Wei Huang, Gaojie Jin, Yi Dong, Changshun Wu, Saddek Bensalem, Ronghui Mu, Yi Qi, Xingyu Zhao, Kaiwen Cai, Yanghao Zhang, Sihao Wu, Peipei Xu, Dengyu Wu, Andre Freitas, and Mustafa A. Mustafa. 2023. [A survey of safety and trustworthiness of large language models through the lens of verification and validation](#). *Preprint*, arXiv:2305.11391.
- Eyke Hüllermeier and Willem Waegeman. 2021. [Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods](#). *Machine Learning*, 110(3):457–506.
- Tomoharu Iwata and Zoubin Ghahramani. 2017. [Improving output uncertainty estimation and generalization in deep learning via neural network gaussian processes](#). *Preprint*, arXiv:1707.05922.
- Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. 1999. An introduction to variational methods for graphical models. *Machine learning*, 37:183–233.
- Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.
- Armen Der Kiureghian and Ove Ditlevsen. 2009. [Aleatory or epistemic? does it matter?](#) *Structural Safety*, 31(2):105–112. Risk Acceptance and Risk Communication.
- Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. [Large language models are zero-shot reasoners](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc.
- Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*.
- S. Kullback and R. A. Leibler. 1951. [On information and sufficiency](#). *The Annals of Mathematical Statistics*, 22(1):79–86.
- Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30.
- Benjamin Lambert, Florence Forbes, Senan Doyle, Harmonie Dehaene, and Michel Dojat. 2024. [Trustworthy clinical ai solutions: A unified review of uncertainty quantification in deep learning models for medical image analysis](#). *Artificial Intelligence in Medicine*, 150:102830.
- Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. 2018a. [A simple unified framework for detecting out-of-distribution samples and adversarial attacks](#). In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.
- Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. 2018b. [A simple unified framework for detecting out-of-distribution samples and adversarial attacks](#). *Preprint*, arXiv:1807.03888.
- Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. [HaluEval: A large-scale hallucination evaluation benchmark for large language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464, Singapore. Association for Computational Linguistics.
- Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. [Teaching models to express their uncertainty in words](#). *Preprint*, arXiv:2205.14334.

- Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2024. [Generating with confidence: Uncertainty quantification for black-box large language models](#). *Preprint*, arXiv:2305.19187.
- Jeremiah Zhe Liu, Zi Lin, Shreyas Padhy, Dustin Tran, Tania Bedrax-Weiss, and Balaji Lakshminarayanan. 2020. [Simple and principled uncertainty estimation with deterministic deep learning via distance awareness](#). *Preprint*, arXiv:2006.10108.
- Luping Liu, Meiling Wang, Mozhi Zhang, Linbo Qing, and Xiaohai He. 2022. Uamner: uncertainty-aware multimodal named entity recognition in social media posts. *Applied Intelligence*, 52(4):4109–4125.
- Andrey Malinin. 2019. *Uncertainty estimation in deep learning with application to spoken language assessment*. Ph.D. thesis, University of Cambridge.
- Andrey Malinin and Mark Gales. 2018. [Predictive uncertainty estimation via prior networks](#). *Preprint*, arXiv:1802.10501.
- Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. [Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models](#). *Preprint*, arXiv:2303.08896.
- Putra Manggala and Christof Monz. 2023. Aligning predictive uncertainty with clarification questions in grounded dialog. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14988–14998.
- Katerina Margatina, Loic Barrault, and Nikolaos Aletras. 2021. On the importance of effectively adapting pretrained language models for active learning. *arXiv preprint arXiv:2104.08320*.
- Katerina Margatina, Timo Schick, Nikolaos Aletras, and Jane Dwivedi-Yu. 2023. [Active learning principles for in-context learning with large language models](#). *Preprint*, arXiv:2305.14264.
- Maria Mora-Cross and Saul Calderon-Ramirez. 2024. Uncertainty estimation in large language models to support biodiversity conservation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, pages 368–378.
- Jishnu Mukhoti, Andreas Kirsch, Joost van Amersfoort, Philip H.S. Torr, and Yarin Gal. 2023. Deep deterministic uncertainty: A new simple baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24384–24394.
- Jie Ren, Jiaming Luo, Yao Zhao, Kundan Krishna, Mohammad Saleh, Balaji Lakshminarayanan, and Peter J. Liu. 2022. Out-of-distribution detection and selective generation for conditional language models. In *The Eleventh International Conference on Learning Representations*.
- Jie Ren, Jiaming Luo, Yao Zhao, Kundan Krishna, Mohammad Saleh, Balaji Lakshminarayanan, and Peter J. Liu. 2023. [Out-of-distribution detection and selective generation for conditional language models](#). *Preprint*, arXiv:2209.15558.
- Alessandro Renda, Marco Barsacchi, Alessio Bechini, and Francesco Marcelloni. 2019. [Comparing ensemble strategies for deep learning: An application to facial expression recognition](#). *Expert Systems with Applications*, 136:1–11.
- Christopher Schröder, Lydia Müller, Andreas Niekler, and Martin Potthast. 2021. Small-text: Active learning for text classification in python. *arXiv preprint arXiv:2107.10314*.
- Xinyue Shen, Zeyuan Chen, Michael Backes, and Yang Zhang. 2023. [In chatgpt we trust? measuring and characterizing the reliability of chatgpt](#). *Preprint*, arXiv:2304.08979.
- Yanyao Shen, Hyokun Yun, Zachary C Lipton, Yakov Kronrod, and Animashree Anandkumar. 2017. Deep active learning for named entity recognition. *arXiv preprint arXiv:1707.05928*.
- Kumar Shridhar, Felix Laumann, and Marcus Liwicki. 2019. [A comprehensive guide to bayesian convolutional neural network with variational inference](#). *Preprint*, arXiv:1901.02731.
- Vaishnavi Shrivastava, Percy Liang, and Ananya Kumar. 2023. Llamas know what gpts don’t show: Surrogate models for confidence estimation. *arXiv preprint arXiv:2311.08877*.
- A Simpkin and Richard Schwartzstein. 2016. Tolerating uncertainty—the next medical revolution? *New England Journal of Medicine*, 375(18).
- Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.
- Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D. Manning. 2023. [Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback](#). *Preprint*, arXiv:2305.14975.
- Artem Vazhentsev, Gleb Kuzmin, Akim Tsvigun, Alexander Panchenko, Maxim Panov, Mikhail Burtsev, and Artem Shelmanov. 2023. [Hybrid uncertainty quantification for selective text classification in ambiguous tasks](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11659–11681, Toronto, Canada. Association for Computational Linguistics.

- Haoyu Wang, Hongming Zhang, Yuqian Deng, Jacob R. Gardner, Dan Roth, and Muhao Chen. 2022. [Extracting or guessing? improving faithfulness of event temporal relation extraction](#). *Preprint*, arXiv:2210.04992.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. [Self-consistency improves chain of thought reasoning in language models](#). *Preprint*, arXiv:2203.11171.
- Yeming Wen, Dustin Tran, and Jimmy Ba. 2020. [Batchensemble: An alternative approach to efficient ensemble and lifelong learning](#). *Preprint*, arXiv:2002.06715.
- Lisa Wimmer, Yusuf Sale, Paul Hofman, Bernd Bischl, and Eyke Hüllermeier. 2023. [Quantifying aleatoric and epistemic uncertainty in machine learning: Are conditional entropy and mutual information appropriate measures?](#) In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pages 2282–2292. PMLR.
- Yijun Xiao and William Yang Wang. 2018. [Quantifying uncertainties in natural language processing tasks](#). *Preprint*, arXiv:1811.07253.
- Yijun Xiao and William Yang Wang. 2021. [On hallucination and predictive uncertainty in conditional language generation](#). *Preprint*, arXiv:2103.15025.
- Yuxin Xiao, Paul Pu Liang, Umang Bhatt, Willie Neiswanger, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2022. [Uncertainty quantification with pre-trained language models: A large-scale empirical analysis](#). In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7273–7284, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2024. [Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms](#). *Preprint*, arXiv:2306.13063.
- Hongshen Xu, Zichen Zhu, Situo Zhang, Da Ma, Shuai Fan, Lu Chen, and Kai Yu. 2024. [Rejection improves reliability: Training llms to refuse unknown questions using rl from knowledge feedback](#). *Preprint*, arXiv:2403.18349.
- Gal Yona, Roei Aharoni, and Mor Geva. 2024. [Narrowing the knowledge evaluation gap: Open-domain question answering with multi-granularity answers](#). *arXiv preprint arXiv:2401.04695*.
- Jiaxin Zhang, Zhuohang Li, Kamalika Das, Bradley Malin, and Sricharan Kumar. 2023a. [SAC<sup>3</sup>: Reliable hallucination detection in black-box language models via semantic-aware cross-check consistency](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15445–15458, Singapore. Association for Computational Linguistics.
- Jiaxin Zhang, Zhuohang Li, Kamalika Das, Bradley A. Malin, and Sricharan Kumar. 2024a. [Sac3: Reliable hallucination detection in black-box language models via semantic-aware cross-check consistency](#). *Preprint*, arXiv:2311.01740.
- Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023b. [Siren’s song in the ai ocean: A survey on hallucination in large language models](#). *Preprint*, arXiv:2309.01219.
- Zhen Zhang, Yuhua Zhao, Hang Gao, and Mengting Hu. 2024b. [Linkner: Linking local named entity recognition models to large language models using uncertainty](#). In *Proceedings of the ACM Web Conference 2024*, WWW ’24, page 4047–4058, New York, NY, USA. Association for Computing Machinery.
- Yu Zhu, Yingchun Ye, Mengyang Li, Ji Zhang, and Ou Wu. 2023. [Investigating annotation noise for named entity recognition](#). *Neural Computing and Applications*, 35(1):993–1007.