Uncertainty Quantification and Confidence Calibration in Large Language Models: A Survey

Xiaoou Liu*
xiaoouli@asu.edu
Arizona State University
Tempe, Arizona, USA

Tiejin Chen*
tchen169@asu.edu
Arizona State University
Tempe, Arizona, USA

Longchao Da
longchao@asu.edu
Arizona State University
Tempe, Arizona, USA

Chacha Chen
chacha@uchicago.edu
University of Chicago
Chicago, IL, USA

Zhen Lin
zhenlin4@illinois.edu
University of Illinois
Urbana-Champaign
Champaign, IL, USA

Hua Wei†
hua.wei@asu.edu
Arizona State University
Tempe, Arizona, USA

Abstract

Uncertainty quantification (UQ) enhances the reliability of Large Language Models (LLMs) by estimating confidence in outputs, enabling risk mitigation and selective prediction. However, traditional UQ methods struggle with LLMs due to computational constraints and decoding inconsistencies. Moreover, LLMs introduce unique uncertainty sources, such as input ambiguity, reasoning path divergence, and decoding stochasticity, that extend beyond classical aleatoric and epistemic uncertainty. To address this, we introduce a new taxonomy that categorizes UQ methods based on computational efficiency and uncertainty dimensions, including input, reasoning, parameter, and prediction uncertainty. We evaluate existing techniques, summarize existing benchmarks and metrics for UQ, assess their real-world applicability, and identify open challenges, emphasizing the need for scalable, interpretable, and robust UQ approaches to enhance LLM reliability.

CCS Concepts

• **Computing methodologies → Machine learning**; **Natural language processing**.

Keywords

Uncertainty Quantification; Large Language Models

*Both authors contributed equally to this research.
†Corresponding author

1 Introduction

Large Language Models (LLMs) like GPT-4 [1] have achieved remarkable capabilities in text generation, reasoning, and decision-making, driving their adoption in high-stakes domains such as healthcare diagnostics [20, 91], legal analysis [10, 59], and transportation systems [16, 55, 118]. However, their reliability remains a critical concern: LLMs often produce plausible but incorrect or inconsistent outputs, with studies showing that over 30% of answers in medical QA tasks contain factual errors [47]. In sensitive applications, these limitations pose risks ranging from misinformation to life-threatening misdiagnoses, underscoring the urgent need for robust reliability frameworks.

Uncertainty quantification (UQ) emerges as an important mechanism to enhance LLM reliability by explicitly modeling confidence in model outputs. By estimating uncertainty, users can identify low-confidence predictions for human verification, prioritize high-certainty responses, and mitigate risks like overconfidence in hallucinations [71]. For instance, in clinical settings, uncertainty-aware LLMs could flag uncertain diagnoses for specialist review, reducing diagnostic errors by up to 41% [99]. This capability is particularly critical as LLMs' transition from experimental tools to production systems requiring accountability.

Traditional UQ methods face significant hurdles when applied to LLMs. Bayesian approaches like Monte Carlo dropout [28] are computationally prohibitive for trillion-parameter models and natural language generation (NLG) tasks, while ensemble methods struggle with consistency across diverse decoding strategies [73]. Furthermore, LLMs introduce unique uncertainty sources, such as input ambiguity [7, 33], reasoning path divergence, and decoding stochasticity that transcend classical aleatoric and epistemic categorizations [42]. The complexity of LLMs, characterized by sequence generation over vast parameter spaces and reliance on massive datasets, exacerbates uncertainty challenges. This complexity, coupled with the critical need for reliable outputs in high-stakes applications, positions UQ for LLMs as a compelling yet underexplored research frontier.

Targeting the unique challenges of UQ in LLMs, this survey firstly introduces a novel taxonomy for LLM UQ, categorizing methods along two axes: (1) computational efficiency (e.g., single-pass vs. sampling-based techniques) and (2) uncertainty dimensions (input, reasoning, parametric, predictive). This framework addresses three gaps in prior works: First, it decouples uncertainty sources unique to LLMs from traditional ML contexts. Second, it evaluates methods through the lens of different dimensions of the responses from LLM: input uncertainty, reasoning uncertainty, parameter uncertainty, and prediction uncertainty. Each of these dimensions may involve aleatoric uncertainty, epistemic uncertainty, or a mixture of both. Third, it identifies understudied areas like reasoning uncertainty, challenges, and possible future directions.

Connection to Existing Surveys: Prior surveys [37, 40, 102] focus on hallucination detection or retrofitting classical UQ taxonomies, neglecting LLM-specific challenges like prompt-driven input uncertainty. Our work uniquely addresses the interplay between model scale, open-ended generation, and uncertainty dynamics, which are critical for modern LLMs but overlooked in earlier frameworks.

The remainder of this survey is structured as follows: Section 2 characterizes LLM uncertainty dimensions and differentiates confidence from uncertainty. Section 3 evaluates UQ methods using our taxonomy. Section 4 introduces the evaluation of UQ methods for LLM, including benchmarks and metrics. Sections 5 and 6 introduce the applications of UQ in different domains with LLMs and identify open challenges and future directions.

2 Perliminaries

2.1 Sources of Uncertainty in LLMs

2.1.1 Aleatoric vs. Epistemic Uncertainty. For UQ on traditional machine learning tasks such as classification or regression [129], there are mainly two types of uncertainty [23, 127]: aleatoric uncertainty, which models the uncertainty from noise in the dataset, and epistemic uncertainty, which arises from the model's lack of knowledge about the underlying data distribution.

Aleatoric uncertainty in LLMs primarily stems from data sources used to train LLMs, which contain inconsistencies, biases, and contradicting information. Additionally, ambiguity in natural language also contributes to aleatoric uncertainty, as different interpretations of the same prompt can lead to multiple plausible responses. On the other hand, when encountering unfamiliar topics, LLMs may exhibit high epistemic uncertainty, often manifesting as hallucinations or overconfident yet incorrect statements. Epistemic uncertainty can be reduced through domain-specific fine-tuning or retrieval-augmented generation techniques that allow the model to access external knowledge sources.

2.1.2 Uncertainty with Different Dimensions. While the uncertainty for LLMs can also be classified through aleatoric and epistemic uncertainty, these two categories alone are insufficient to fully capture the complexities of uncertainty in LLMs. In particular, LLMs exhibit uncertainty not only due to training data limitations but also due to input variability and decoding mechanisms. Therefore, in the following, we formulate four dimensions of uncertainty, each of which may involve aleatoric uncertainty, epistemic uncertainty,

or a combination of both. The technical methods on how to quantify these uncertainties will be discussed later in Section 3.

- **Input Uncertainty** (Aleatoric Uncertainty): Input uncertainty arises when a prompt is ambiguous or underspecified, making it impossible for an LLM to generate a single definitive response. This is inherently aleatoric, as even a "perfect model" cannot resolve the ambiguity. For instance, *"What is the capital of this country?"* lacks sufficient context, leading to unpredictable outputs. Similarly, *"Summarize this document"* may yield different responses depending on different expected details.

- **Reasoning Uncertainty** (Mixed Uncertainty): Reasoning uncertainty occurs when an LLM derives answers through multi-step reasoning [81] or retrieval [61], where the uncertainty of each step can lead to ambiguous or incorrect results. This uncertainty is aleatoric when the problem itself is ambiguous and epistemic when the model cannot offer robust reasoning.

- **Parameter Uncertainty** (Epistemic Uncertainty): Parameter uncertainty stems from training data gaps, where the model has either never seen relevant information or has learned an incorrect representation. Unlike aleatoric uncertainty, epistemic uncertainty can be reduced by improving the model's knowledge base. Bayesian methods [28], deep ensembles [56], and uncertainty-aware training [83] can help quantify and mitigate this type of uncertainty.

- **Prediction Uncertainty** (Mixed Uncertainty): Prediction uncertainty refers to variability in generated outputs across different sampling runs, influenced by both aleatoric and epistemic sources. For example, when asked *"What are the side effects of a new experimental drug?"*, the model's responses might vary significantly across different sampling runs, especially if no reliable data is available in its training set. A high-variance output distribution in such scenarios suggests that the model is both aware of multiple possible answers, reflecting aleatoric uncertainty, and uncertain due to incomplete knowledge, highlighting epistemic uncertainty.

2.2 Uncertainty and Confidence in LLMs

2.2.1 Classical Confidence Estimation. UQ and confidence estimation are closely related yet distinct concepts. In traditional machine learning, uncertainty is a property of the model's predictive distribution, capturing the degree of variability or unpredictability *given a particular input*. In contrast, confidence reflects the model's belief in the correctness of a particular prediction. If we follow the definition in classification tasks, the confidence measure would be the predicted probability $\hat{p}(Y = y|x)$ given input x (an uncertainty measure which does not depend on the particular prediction y could be entropy, taking the form of $\sum_y -\hat{p}(Y = y|x) \log \hat{p}(Y = y|x)$). Table 1 shows a similar notation in a question-answering (QA) task in Natural Language Generation (NLG). The corresponding **confidence score** in NLG tasks for an auto-regressive language model would be the joint probability for the generated sequence:

$$C(\mathbf{x}, \mathbf{s}) = \hat{p}(\mathbf{s}|\mathbf{x}) = \prod_i \hat{p}(s_i|\mathbf{s}_{<i}, \mathbf{x}). \tag{1}$$

The log of Eq. (1) is sometimes referred to as *sequence likelihood* [139]. In general, an uncertainty estimate in existing literature usually takes the form of $U(\mathbf{x})$, while confidence estimates are usually expressed as $C(\mathbf{x}, \mathbf{s})$. Note that, unlike classification tasks, not

Notation	Description
$\mathbf{x}$	The question that LLMs answer
$\mathbf{s}$	Generation from LLMs
w_i	i-th token in the generation $\mathbf{s}$
$\mathcal{D}$	Corpus of LLMs
$U(\mathbf{x})$	Uncertainty of question $\mathbf{x}$
$C(\mathbf{x}, \mathbf{s})$	Confidence of generation $\mathbf{s}$ given $\mathbf{x}$
$H(\mathbf{s})$	Entropy of generation $\mathbf{s}$

Table 1: Notations used in this paper for an exemplary QA task.

all NLG applications have the notion of a "correct" answer (e.g., summarization). Thus, while for the ease of writing we use the term *correctness* throughout this section, it should really be interpreted as the gold-label for the particular application. Note also that in most cases, the correct answer is not unique, and thus such gold-label typically takes the form of a "correctness function" that decides whether a particular generation $\mathbf{s}$ is good or not. We will denote such a function as $f(\mathbf{s}|\mathbf{x})$.

2.2.2 Confidence Improvement. There are usually two dimensions along which researchers improve confidence estimates in NLG, which is unsurprisingly largely influenced by confidence scoring literature from classification [45], especially binary classification. We refer to them as *ranking performance* and *calibration*:

• **Ranking performance** refers to the discriminative power of the confidence measure on the correctness. Like in classification, LLM confidence is often evaluated by its ability to separate correct and incorrect answers, thus typically measured by evaluation metrics like AUROC [49] or AUARC [67] as detailed in Section 4.

• **Calibration** refers to closing the gap between the confidence score and the expected correctness *conditioned on confidence score*. It has a long history preceding even modern machine learning [85], but bears slightly different meanings in NLP. In general, we could define a perfectly calibrated confidence measure to achieve: $\forall c, \mathbb{E}[f(\mathbf{s}|\mathbf{x})|C(\mathbf{x}, \mathbf{s}) = c] = c$, where the expectation is taken over the joint distribution of $\mathbf{x}$ and generation $\mathbf{s}$. A lot of papers focus on evaluating the calibration quality of specific language models (LMs) and tasks [52, 114]. Evaluation typically relies on variants of Expected Calibration Error (ECE) [52, 107]. Oftentimes confidence scores from classification could be directly applied [103] in order to evaluate whether an LM is over- or under-confident, especially for de facto classification tasks like sentiment analysis or multiple-choice QA.

2.2.3 Confidence Estimation for LLMs. Confidence estimation in large language models (LLMs) refers to the task of quantifying how certain a model is about a specific generated output. In this subsection, we review three major families of approaches to confidence estimation in LLMs:

• **UQ methods with Confidence Estimation.** As uncertainty and confidence are often intertwined, many approaches used in UQ have their counterpart in confidence estimation. For example, for black-box settings where the parameters of LLMs are unavailable, [66, 133] computes a similarity matrix of sampled responses and derives confidence estimates for each generation via its degree or distance derived from the graph Laplacian, before using these scores to compute uncertainty. For white-box settings where model

parameters are available, researchers mostly compute the confidence from the output logits, either through normalizing Eq. (1) with the length of $\mathbf{s}$ [75], replacing the logit-sum or mean with weighted sum by attention values [67] or by importance inferred from natural language inference (NLI) models [24]. Such variants of sequence likelihood could then be fed for (entropy-style) uncertainty computation [51, 67].

• **LLM-as-a-judge.** Another popular approach is asking the LM itself whether a particular free-form generation is correct [49]. However, this formulation also poses a restriction on the confidence estimation method, as it is essentially a scalar logit. Thus, many extensions focus on applying calibration methods from classification to calibrate such self-evaluation. The few exceptions include [49, 97], which converts samples from free-form generation into a *multiple-choice* question (with generations being the options) and adds a "None of the above" option to elicit the confidence.

• **Trainable Confidence Estimators.** Since we typically care about the LM's confidence in the "semantic space" due to semantic invariance, instead of manipulating logits, a popular approach is to perform additional training for confidence estimation. This could be done on the base LM (either fully [46, 50, 139] or partially [71]) with a different loss, or using a separate model on the internal or external representations from the base LM [44, 109]. On the other end of the spectrum, without any training, prompting could be used to elicit verbalized confidence values [107]. Finally, one could combine multiple confidence estimation methods and enjoy the benefit of ensembling [29].

As with UQ evaluation (more in Section 4), the choice of correctness function has a profound impact on the conclusion of the experiments, especially for free-form generation tasks. Popular choices include using (potentially larger) LLM as judges [66, 71, 107], human annotations [97], or lexical similarities such as ROUGE [51, 139]. Recently, Liu et al. [72] proposes to evaluate free-form generation confidence measures with selected multiple-choice datasets as an efficient complement. For longer generations, Huang et al. [41] proposes to use ordinal (not binary) correctness values to capture the ambiguity in the quality of a generation. In a similar flavor, [3] studies the issues in the evaluation of calibration when there is intrinsic human disagreement on the label.

Remarks. Existing literature sometimes uses the terms uncertainty and confidence interchangeably. They do often seemingly coincide: When a model's prediction has low confidence, we naturally consider this as a high uncertainty case. This, however, is treating $U(\mathbf{x}) = -\max_{\mathbf{s}} C(\mathbf{x}, \mathbf{s})$ as an uncertainty estimate. In general, a model may exhibit high uncertainty over its output space but still express high confidence in a specific output. Conversely, a model could have low overall uncertainty but low confidence in a particular prediction. While the "low uncertainty low confidence case" is relatively less interesting in classification or regression tasks due to MLE point prediction, this scenario is notably more common in NLG, as the output is typically *randomly sampled*[1] from the predictive distribution. There are also applications that require one but not the other (e.g. conformal language modeling [92] or *seletive generation* [13]). In the rest of this paper, we sometimes

---

[1]In fact, even if the output is greedily generated, it might not have the highest confidence as measured by Eq. (1).

Method	Uncertainty Dimensions	Efficency Features	Access to Model	Confidence
Input clarification ensembles [36]	Input Uncertainty	Multi Rounds Generations	Black-box	No
ICL-Sample [68]	Input Uncertainty	Multi Rounds Generations	Black-box	No
SPUQ [29]	Input Uncertainty	Multi Rounds Generations + Additional Model	Black-box	No
UAG [128]	Reasoning Uncertainty	Single Round Generation	White-box	No
CoT-UQ [132]	Reasoning Uncertainty	Single Round Generation	White-box	Yes
TouT [80]	Reasoning Uncertainty	Multi Rounds Generations	Black-box	No
TopologyUQ [18]	Reasoning Uncertainty	Multi Rounds Generations	Black-box	No
Stable Explanations Confidence [5]	Reasoning Uncertainty	Multi Rounds Generation	Black-box	Yes
SAPLMA [2]	Parameter + Prediction Uncertainty	Fine-tuning	White-box	Yes
Supervised estimation[69]	Parameter + Prediction Uncertainty	Fine-tuning	White-box	Yes
UaIT [70]	Parameter + Prediction Uncertainty	Fine-tuning	White-box	Yes
LoRA ensembles [4]	Parameter Uncertainty	Fine-tuning	White-box	Yes
BloB [115]	Parameter Uncertainty	Fine-tuning	White-box	Yes
BLoRA [121]	Parameter Uncertainty	Fine-tuning	White-box	Yes
Perplexity [77, 82]	Prediction Uncertainty	Single Round Generation	White-box	Yes
SAR [24]	Prediction Uncertainty	Single Round Generation	White-box	Yes
P(True) [49]	Prediction Uncertainty	Single Round Generation	White-box	Yes
Response improbability [26]	Prediction Uncertainty	Single Round Generation	White-box	Yes
Average log probability [76]	Prediction Uncertainty	Single Round Generation	White-box	Yes
Predictive Entropy [49]	Prediction Uncertainty	Multi Rounds Generations	White-box	Yes
Relative Mahalanobis distance [96]	Prediction Uncertainty	Multi Rounds Generations	White-box	Yes
HUQ [112]	Prediction Uncertainty	Multi Rounds Generations	White-box	Yes
Conformal Prediction (CP) [53, 92]	Prediction Uncertainty	Multi Rounds Generations	White-box	No
ConU [116]	Prediction Uncertainty	Multi Rounds Generations	White-box	No
Level-adaptive CP [11]	Prediction Uncertainty	Multi Rounds Generations	White-box	No
LoFreeCP [104]	Prediction Uncertainty	Multi Rounds Generations	Black-box	No
Ecc(J),Deg(J) [66]	Prediction Uncertainty	Multi Rounds Generations	Black-box	Yes
Eig(J) [66]	Prediction Uncertainty	Multi Rounds Generations	Black-box	No
Normal length predictive entropy [75]	Prediction Uncertainty	Multi Rounds Generations +Additional Model	White-box	Yes
Semantic Entropy [51]	Prediction Uncertainty	Multi Rounds Generations + Additional Model	White-box	Yes
Kernel Language Entropy [87]	Prediction Uncertainty	Multi Rounds Generations + Additional Model	White-box	Yes
Ecc(C),Ecc(E),Deg(C),Deg(E) [66]	Prediction Uncertainty	Multi Rounds Generations + Additional Model	Black-box	Yes
Eig(C),Eig(E) [66]	Prediction Uncertainty	Multi Rounds Generations + Additional Model	Black-box	No
MD-UQ [8]	Prediction Uncertainty	Multi Rounds Generations + Additional Model	Black-box	No
D-UE [15]	Prediction Uncertainty	Multi Rounds Generations + Additional Model	Black-box	Yes

Table 2: An overview of UQ methods discussed in this paper for different dimensions, efficiency, and model settings.

follow the language of the original papers and treat confidence estimates as uncertainty, but will clearly mark the methods that provide confidence estimates.

3 UQ Methods for Different Dimensions

3.1 Input Uncertainty

As mentioned in Section 2.1.2, input uncertainty arises from the ambiguous or incomplete input to the LLMs. While there are works in the LLMs domain that try to benchmark or deal with ambiguity [7, 22, 33, 130], they did not model the uncertainties induced by ambiguity. Existing UQ methods that specifically consider input uncertainty focus on perturbing the input prompts of LLMs. For instance, [36] proposes an approach that generates multiple clarifications for a given prompt and ensembles the resulting generations by using mutual information to capture the disagreement among the predictions arising from different clarifications. Similarly, [68] proposed ICL-Sample, which quantified the input uncertainty in the setting of in-context learning using different in-context samples. [29] proposes SPUQ, which perturbs the input by techniques such as paraphrasing and dummy tokens to expose the model's sensitivity and capture uncertainty. Specifically, SPUQ quantified the input uncertainty by using a similarity metric such as BERTScore [135] to measure how consistent the responses are across different perturbations. In general, there are only a few papers that consider input uncertainty. Since ambiguity is common and important in natural language, more effort is needed into input uncertainty and its application.

3.2 Reasoning Uncertainty

Reasoning is the process of drawing conclusions based on available information. As the LLMs have demonstrated remarkable performance on tasks involving reasoning, recent research has focused on using UQ in LLM reasoning and analyzing the internal reasoning process. For example, TopologyUQ [18] introduces a formal method to extract and structure LLM explanations into graph representations, quantifying reasoning uncertainty by employing graph-edit distances and revealing redundancy through stable topology measures. Stable-Explanation Confidence [5] treats each possible model and its explanation pair as a test-time classifier to construct a posterior answer distribution that reflects overall reasoning confidence. CoT-UQ [132] integrates chain-of-thought reasoning into a response-level UQ framework, thereby leveraging the inherent multi-step reasoning capability of LLMs to further improve uncertainty assessment. Collectively, these approaches provide a robust and interpretable framework for enhancing LLM reasoning by quantifying uncertainty at local or global levels.

The quantified uncertainty could be used to guide the exploration of reasoning steps and improving the final performance in completing the tasks. In [80], they propose Tree of Uncertain Thoughts (TouT), which extend the Tree of Thoughts (ToT) [125] framework

by quantifying the uncertainties in intermediate reasoning steps with Monte Carlo Dropout and assigning uncertainty scores to important decision points. Similarly, [128] reduces the error accumulation in multi-step reasoning by monitoring the predicted probability of the next token at each generation step, dynamically retracting to more reliable states and incorporating certified reasoning clues when high uncertainty is detected. Their experimental results shows that integrating uncertainty enhances the precision of generated responses by integrating these local measures with global search techniques.

3.3 Parameter Uncertainty

Parameter uncertainty arises when an LLM lacks sufficient knowledge due to limitations in its training data or model parameters. It reflects the model's uncertainty about its own predictions, which can be reduced with additional training or adaptation techniques.

Traditional UQ methods like Monte Carlo Dropout and Deep Ensembles have been widely used but are computationally infeasible for large-scale LLMs due to the need for multiple forward passes or model replicas. To address this, Bayesian Low-Rank Adaptation by Backpropagation (BLoB) [115] and Bayesian Low-Rank Adaptation (BLoRA) [121] incorporate Bayesian modeling into LoRA adapters, allowing uncertainty estimation through parameter distributions without a full-model ensemble. However, these methods still incur significant computational costs.

Finetuning-based approaches offer a more practical alternative. Techniques such as Supervised Uncertainty Estimation [69] train auxiliary models to predict the confidence of LLM outputs based on activation patterns and logit distributions. Similarly, Uncertainty-aware Instruction Tuning (UaIT) [70] modifies the fine-tuning process to explicitly train models to express uncertainty in their outputs. SAPLMA [2] refines probabilistic alignment techniques to dynamically adjust model uncertainty estimates, ensuring adaptability to different downstream tasks. Additionally, LoRA ensembles [4] provide an alternative to full-model ensembles by training multiple lightweight LoRA-adapted variants of an LLM instead of retraining the entire network.

3.4 Prediction Uncertainty

Most off-the-shelf UQ methods focus on prediction uncertainty since it is the most straightforward way to estimate the uncertainty. Considering the number of generations and models when estimating uncertainties, existing methods for predicting uncertainty can be categorized into the following three categories.

3.4.1 Single Round Generation. Most single-round generation methods utilize the logit or hidden states during generation. With only one round of generation, these methods usually methods are usually efficient in estimating uncertainties.

• **Perplexity** is a measure of how well a probabilistic language model predicts a sequence of text [111] while Mora-Cross and Calderon-Ramirez [82], Margatina et al. [77] and Manakul et al. [76] utilize the perplexity as the uncertainty. In detail, using w_i as the i-th token in the generation, perplexity is given by Perplexity $= \exp\left(-\frac{1}{N}\sum_{i=1}^{N} \ln p(w_i)\right)$. A higher perplexity means

the model spreads its probability more broadly over possible words, indicating that it has a higher uncertainty.

• **Maximum Token Log-Probability.** Apart from the perplexity, Maximum token log-probability [76] measures the sentence's likelihood by assessing the least likely token in the sentence. A higher Maximum(p) indicates higher uncertainty of the whole generation. It is calculated by $Max(p) = \max_i(-\ln p(w_i))$.

• **Entropy** reflects how widely distributed a model's predictions are for a given input, indicating the level of uncertainty in its outputs [49, 51]. Entropy for the i-th token is provided by $\mathcal{H}_i = -\sum_{\tilde{w} \in \mathcal{D}} p_i(\tilde{w}) \log p_i(\tilde{w})$. Then it is possible to use the mean or maximum value of entropy as the final uncertainty [76]: $Avg(\mathcal{H}) = \frac{1}{N}\sum_{i=1}^{N} \mathcal{H}_i; Max(\mathcal{H}) = \max_i(\mathcal{H}_i)$. Furthermore, Shifting Attention to Relevance (SAR) [24], enhanced the performance of entropy by adjusting attention to more relevant tokens inside the sentence. In detail, SAR assigned weight for $\mathcal{H}_i$ and the weight $R(w_i, s, x)$ can be obtained by: $R(w_i, s, x) = 1 - |g(x \cup s, x \cup s \setminus \{w_i\})|$, where g is a function that measures the semantic similarity between two sentences, which can be estimated with NLI models [24].

• **Response Improbability** [26] uses response improbability, which computes the probability of a given sentence and subtracts the resulting value from one. In detail, response improbability is provided by $MP(s) = 1 - \prod_{i=1} p_i(w_i)$. If the sentence is certain (i.e., the product of token probabilities is high), $MP(s)$ will be low.

• **P(True)** [49] measures the uncertainty of the claim by asking the LLM itself whether the generation is true or not. Specifically, P(True) is calculated [2]: P(True) $= 1 - p(y_1 = $ "True"$)$. Note that here we are using y_1 as the first token instead of w_1 because w_1 represents the first token in the generation s while y_1 represents the first token when asking LLM whether the generation s is correct or not. P(True) requires running the LLM twice. However, it does not require multiple generations s. Therefore, we still classify this method as a single-round generation [3].

3.4.2 Multiple rounds generation. Multiple rounds generation methods estimate uncertainty by generating multiple predictions from the LLMs and analyzing their consistency, similarity, or variability. These approaches assume that if a model is confident, its outputs should be stable across different sampling conditions.

• **Token-Level Entropy.** Token-level entropy quantifies uncertainty in LLMs by analyzing the probability distribution of generated tokens across multiple samples. A confident model assigns high probability to a specific token, resulting in low entropy, while uncertain predictions distribute probability across multiple tokens, leading to higher entropy.

Multiple responses are generated for the same input to estimate token-level entropy, and the entropy of the token probability distribution is computed. For example, predictive entropy [49] can also be applied to multiple response settings and shows a better uncertainty quality based on the variability of multiple outputs. Similarly, SAR [24] could also be applied to multiple responses. [75] extends with Monte Carlo-based approximations and focuses on

---

[2]The original name is P(IK), which stands for *"I Know"*.
[3]This could be considered an uncertainty estimate as the sequence to be evaluated is the prediction given the input.

how probability distributions evolve across tokens during autoregressive generation. There are two main approaches to get the final uncertainty: one averages entropy across multiple sampled outputs, and the other decomposes sequence-level uncertainty into token-level contributions using entropy approximation.

• **Conformal Prediction.** Conformal Prediction (CP) [101] is a statistical framework that provides formal coverage guarantees for uncertainty estimation in LLMs. Its distribution-free properties make it suitable for both black-box and white-box models.

In the black-box setting, where model internals are inaccessible, CP estimates uncertainty using response frequency, semantic similarity, or self-consistency. One study proposes a method tailored for API-only LLMs [104], using frequency-based sampling combined with normalized entropy and semantic similarity to define nonconformity scores. Another black-box CP method introduces a self-consistency-based uncertainty measure [116], which clusters sampled generations and selects a representative response to construct prediction sets with correctness guarantees, making it particularly effective for open-ended NLG tasks.

On the other hand, white-box CP methods use logits, internal activations, and calibration techniques for more refined uncertainty estimation. One study proposes Conformal Language Modeling [92], which integrates CP with autoregressive text generation by dynamically calibrating a stopping rule to ensure at least one response in the generated set is statistically valid. Another work adapts CP for multiple-choice QA [53], using model confidence scores to calibrate prediction sets, ensuring coverage with minimal set size. A more advanced technique, conditional CP [11], dynamically adjusts coverage guarantees based on the difficulty of the input, optimizing prediction set size while maintaining reliability.

• **Consistency-Based Methods.** Consistency-based uncertainty estimation methods analyze the agreement between multiple generated responses from an LLM to determine uncertainty. The underlying assumption is that if the model is confident, its responses should be consistent, while high variability among responses suggests uncertainty. [66] measures the overlap between words through Jaccard similarity in different generations. This method evaluates the deviation from self-consistency, where a high Jaccard similarity across generations implies low uncertainty.

However, word-level similarity alone is insufficient, as different responses can convey the same meaning using different phrasing. Moreover, the generated response might include long reasoning steps that require detailed analysis [32]. To address this problem, some methods incorporate external models to assess semantic similarity rather than relying solely on lexical overlap.

3.4.3 Multiple Rounds Generation with External Models. Semantic-based uncertainty estimation methods expand multiple generation approaches by incorporating external models, such as Natural Language Inference (NLI) or pretrained language models, to evaluate the semantic relationships among generated responses beyond surface-level similarity.

• **Distribution-based entropy methods** quantify uncertainty by modeling the distribution of generated responses in a semantic space. Semantic Entropy (SE) [51] refines uncertainty estimation by clustering generated responses based on semantic equivalence.

This approach uses an NLI model to determine entailment relationships among responses, grouping them into meaning-preserving clusters. Instead of calculating entropy over individual responses, SE computes entropy over these clusters. Kernel Language Entropy (KLE) [87] takes a different approach by avoiding explicit clustering. Instead, it embeds the responses in a semantic space using a positive semidefinite kernel function. By computing von Neumann entropy over these response distributions, KLE provides an even more fine-grained measure of uncertainty that considers nuanced semantic variations.

• **Pairwise similarity methods** construct a pairwise semantic similarity matrix between responses and analyze its structural properties to estimate uncertainty. Methods like [66] use NLI models to score entailment and contradiction between every pair of generated outputs, forming a weighted similarity graph. A confident model yields semantically coherent responses with strong mutual agreement (high similarity), while inconsistent or ambiguous outputs lead to greater dispersion in the matrix. To quantify this dispersion, spectral graph metrics are applied: Eccentricity (Ecc) captures variability spread, Eigenvalue-based (Eig) measures assess global structure, and Degree (Deg) evaluates local consistency. Recent works further extend this by modeling the response similarity graph as directed [15] or multi-dimensional [8], allowing for richer representation of semantic asymmetry or latent factors in uncertainty.

4 Evaluation of Uncertainty in LLMs

4.1 Benchmark Datasets

Datasets used in previous studies can be organized into several categories based on their focus. An overall summary of the categorization of datasets and benchmarks for UQ is shown in Table 3.

• **Reading comprehension** benchmarks include CoQA [94] for conversational question answering tasks, RACE [54] for general reading comprehension, TriviaQA [48] for fact-based questions, CosmosQA [39] for contextual understanding, SQuAD [93] for question answering on passages, and HotpotQA [123] for multi-hop reasoning.

• **Reasoning and math** benchmarks include HotpotQA [123] and StrategyQA [30], which test multi-hop reasoning, GSM8K [12] for solving math problems, and CalibratedMath [64], designed to evaluate confidence expression in arithmetic. These benchmarks are helpful to evaluate the reasoning uncertainty.

• **Factuality** evaluation draws on datasets such as TruthfulQA [65] for addressing common misconceptions, FEVER [106] for claim verification, HaluEval [58] for detecting hallucinations, and an annotated FActScore [78] dataset for evaluating the factuality of long-form text generated by LLMs.

• **General knowledge** benchmarks can be adapted for UQ to test the models' general knowledge, such as MMLU [34] for a wide range of subjects, GPQA [95] for multiple-choice questions in physical sciences, and HellaSwag [131] for common-sense reasoning through sentence completion. These benchmarks can be adapted for UQ because the tasks can be reduced to a classification problem, determining whether the model is confident or uncertain. The structured nature of these benchmarks allows for clear evaluation of the model's confidence in its predictions.

Category	Benchmarks
Reading Comprehension	TriviaQA [48], CoQA [94], RACE [54], CosmosQA [39], SQuAD [93], HotpotQA [123]
Reasoning & Math	StrategyQA [30], HotpotQA [123], GSM8K [12], CalibratedMath [64]
Factuality	TruthfulQA [65], FEVER [106], HaluEval [58], FActScore [78]
General Knowledge	MMLU [34], GPQA [95], HellaSwag [131]
Consistency & Ambiguity	ParaRel [25], AmbigQA [79], AmbigInst [36], Abg-SciQA [7]

Table 3: Categorization of benchmarking datasets for UQ.

• **Consistency and ambiguity** are two additional kind of benchmarks for UQ. Consistency benchmarks such as ParaRel [25] tests semantic consistency across 328 paraphrases for 38 relations, and datasets like AmbigQA and AmbigInst, which feature inherent ambiguities [7, 36, 79]. Ambiguity datasets are useful in UQ evaluation because they introduce aleatoric uncertainty by highlighting cases where multiple plausible interpretations exist, helping to assess how well models distinguish between data-driven randomness and model-based uncertainty. These datasets enable a more precise decomposition of uncertainty into aleatoric and epistemic components, improving model reliability and interpretability.

Recently, there have been efforts to develop UQ benchmarks for dedicated sources of uncertainty or specific methods. For example, MAQA [122] is a dataset specifically designed to evaluate epistemic uncertainty in language models; LM-Polygraph [27] was later adopted as a comprehensive uncertainty benchmark [110]. [126] developed a benchmark for conformal prediction methods. These contributions represent specialized datasets explicitly designed to assess UQ capabilities in LLMs, rather than adapting existing general-purpose benchmarks.

4.2 Evaluation Metrics

UQ is often evaluated from binary classification tasks, with the rationale being that high uncertainty should correspond to low expected accuracy. This is typically modeled by assigning a binary label to each response with a correctness function and using the uncertainty estimates to predict the label. AUROC (Area Under the Receiver Operating Characteristic curve), which measures how effectively the uncertainty score separates correct from incorrect responses, is often used. With values ranging from 0 to 1, higher AUROCs indicate better performance. Responses with confidence above the threshold are classified as predicted positives, while those below are treated as predicted negatives. Many prior studies use AUROC to evaluate how well the uncertainty score discriminates correct from incorrect predictions [6, 51, 69, 72, 119]. Similarly, AUPRC (Area Under the Precision-Recall Curve) and AUARC (Area Under the Accuracy-Rejection Curve) [86] also offer further insights into UQ. AUPRC measures how well the uncertainty score separates correct from incorrect responses [68], while AUARC assesses how effectively the uncertainty measure aids in selecting accurate responses by determining which uncertain questions to reject [67].

In the context of NLG where the correctness label is hard to obtain, researchers also compute heuristic-based fuzzy matching metrics such as BLEU [90] and ROUGE [51] between the generated text and the reference output(s) to gauge the quality. However, these metrics often fail to capture semantic fidelity or factual correctness. Consequently, many researchers are increasingly turning

to LLM-as-a-judge evaluations, wherein a large language model (e.g., GPT-4) is prompted to assess text quality or correctness. This approach can capture nuanced aspects like coherence, style, and factuality, but also introduces risks of bias and inconsistency. Human annotation, however, is expensive and is often limited to a small scale [51, 133].

Apart from the binary classification framework, there are also multiple evaluation methods designed for the specific treatment of uncertainty, sometimes qualitative. For example, focusing on decomposing aleatoric and epistemic uncertainty, [36] evaluates only the aleatoric part by using AmbigQA [79], as high ambiguity questions should incur higher aleatoric uncertainty (whereas math questions, for examples, might have lower). The evaluation in [31], on the other hand, is a comparison between the variability of human production (generation) with that of the LM. With an emphasis on UQ for longer generations, [133] compares the uncertainty estimate against FActScore [78], as the "correctness" of a long paragraph could be ill-defined or ambiguous.

5 UQ Applications in LLMs

LLMs are increasingly applied in diverse domains, offering flexibility and reasoning capabilities. However, UQ is crucial for ensuring their reliability, particularly in high-stakes applications. This section will introduce the applications that integrate the UQ of LLMs from some example domains. Many other fields like energy management, operations research, etc., employ LLMs and would require such discussions on the need for UQ as well.

• **Robotics.** LLM-based robotic planning suffers from ambiguity and hallucinations, motivating the need for UQ in the planning loop. For example, closed-loop planners [141] employ an uncertainty-based failure detector to continuously assess and adjust plans in real-time, while non-parametric UQ methods [108] use an efficient querying strategy to improve reliability. [84] integrates action feasibility checks to align the LLM's confidence with real-world constraints, improving success rates from approximately 70% to 80%. Similarly, [88] dynamically adjusts thresholds for alternative paths in adaptive skill selection, achieving higher success rates. [62] develops an introspective planning framework with LLMs self-assess their uncertainty to enhance safety and human-robot collaboration.

• **Transportation.** Preliminary research explores how LLMs can enhance transportation systems [17, 19, 124]. For example, LLM inference has been used to bridge the sim-to-real gap in traffic signal control [16, 19] and smooth mixed-autonomy traffic [124]. However, both cases reveal the potential risk posed by hallucination. A few works have investigated the uncertainty measure while using the LLMs [21], which tries to link the use of VLMs with deep probabilistic programming for UQ while conducting multimodal traffic accident forecasting tasks.

• **Healthcare.** In healthcare, LLMs and VLMs can be good references for diagnosis, but uncertainty is a critical dimension that should be considered together with the generation of more reliable treatment plans [98]. In [9], it quantifies uncertainty in a white-box setting, and reveals that an effective reduction of model uncertainty can be achieved by using the proposed multi-tasking and ensemble methods in EHRs. However, as [117] benchmarks popular uncertain quantification methods with different model sizes

on medical question-answering datasets, the challenge of UQ for medical applications is still severe.

6 Challenges and Future Directions

While significant strides have been made in integrating uncertainty quantification into LLMs, several unaddressed challenges persist. This section will explore these unresolved issues, ranging from efficiency-performance trade-offs to cross-modal uncertainty, and outline promising avenues for future research, aiming to advance the reliability of LLMs in high-stakes applications.

● **Efficiency-Performance Trade-offs**. Multi-sample uncertainty methods incur prohibitive costs for trillion-parameter LLMs ($12k per million queries [57]), yet yield marginal reliability gains (≤ 0.02 AUROC improvement [120]). Hybrid approaches combining low-cost proxies (attention variance [35], hidden state clustering [87]) could resolve this by achieving 90% of maximal performance at 10% computational cost. For example, precomputing uncertainty "hotspots" during inference could trigger targeted multi-sampling only for high-risk outputs like medical diagnoses.

● **Interpretability Deficits**. Users struggle to distinguish whether uncertainty stems from ambiguous inputs, knowledge gaps, or decoding stochasticity. Modular architectures that decouple uncertainty estimation layers [38, 100] or employ causal tracing of transformer attention pathways [113] could clarify uncertainty origins. For instance, perturbing model weights [28] might reveal parametric uncertainty in low-resource languages, while input modules flag underspecified terms for clarification.

● **Cross-Modality Uncertainty**. Integrating vision, text, and sensor data introduces misaligned confidence estimates between modalities: LVLMs exhibit 2.4× higher uncertainty in visual vs. textual components [136], causing 63% of errors in multi-modal QA [134]. Dynamic contrastive decoding and uncertainty-aware fusion protocols show promise[43, 105], but require domain-specific adaptations (e.g., aligning radiology images with reports [60, 138]). Future work must develop unified uncertainty embeddings to harmonize modality confidence scales and adversarial training against cross-modal backdoor attacks [63, 137].

● **System-level Uncertainty in Agents and Reasoning.** As LLMs are increasingly deployed as autonomous agents or reasoning engines, the propagation and accumulation of uncertainty across steps becomes critical. Errors in early steps can lead to cascading failures, especially when the model expresses misplaced confidence. However, most existing UQ methods operate at one round of outputs from LLM, lacking mechanisms to capture uncertainty over multi-step reasoning chains or multi-action plans. As studies suggest that LLMs often fail to revise earlier decisions when presented with contradictory information [14], there is a need for temporally-aware uncertainty tracking. Enhancing LLMs with structured memory or model-based planning, or leveraging graph-based representations to trace and revise uncertain steps [74, 125], could possibly provide more reliable behavior.

● **UQ Evaluation.** Evaluating the quality of UQ remains a fundamental challenge. While the binary classification metrics introduced in Section 4.2 are widely used, they are not always suitable: many tasks, especially in NLG, cannot be easily reduced to binary correctness. Even for structured tasks like question answering, determining whether a free-form generation is correct can be nontrivial due to semantic variability and ambiguity. This issue becomes even more pronounced in open-ended tasks. Moreover, LLM-as-a-judge evaluation approaches are themselves subject to systematic biases [65, 89, 140]. In addition, common evaluation metrics such as AUROC and AUARC often fail to capture what might be considered "meaningful" uncertainty. These metrics typically assess a model's ability to distinguish between correct and incorrect outputs, but do not differentiate between confidently wrong responses and those accompanied by an appropriate level of uncertainty.

7 Conclusion

In this survey, we offer a comprehensive overview of uncertainty quantification (UQ) in Large Language Models (LLMs). We first introduce the fundamental concepts relevant to both UQ and LLMs, highlighting the importance of reliability in high-stakes applications. Following this, we propose a detailed taxonomy for characterizing uncertainty dimensions in LLMs, including input, reasoning, parameter, and prediction uncertainty. We systematically introduce existing UQ methods using our novel taxonomy, reviewing their effectiveness across different uncertainty types. Ultimately, we identify and discuss some of the persistent challenges in UQ for LLMs, providing insightful directions for future research. The primary goal of this survey is to promote the integration of UQ techniques into LLM development, motivating both machine learning researchers and practitioners to participate in this rapidly advancing area.

ACKNOWLEDGMENTS

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).
[2] Amos Azaria and Tom Mitchell. 2023. The Internal State of an LLM Knows When It's Lying. (2023), 967–976.
[3] Joris Baan, Wilker Aziz, Barbara Plank, and Raquel Fernández. 2022. Stop Measuring Calibration When Humans Disagree. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. 1892–1915.
[4] Oleksandr Balabanov and Hampus Linander. 2024. Uncertainty quantification in fine-tuned LLMs using LoRA ensembles. *arXiv preprint arXiv:2402.12264* (2024).
[5] Evan Becker and Stefano Soatto. 2024. Cycles of thought: Measuring llm confidence through stable explanations. *arXiv preprint arXiv:2406.03441* (2024).
[6] 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*. 5186–5200.
[7] Tiejin Chen, Kuan-Ru Liou, Mithun Shivakoti, Aaryan Gaur, Pragya Kumari, Meiqi Guo, and Hua Wei. 2025. Abg-SciQA: A dataset for Understanding and Resolving Ambiguity in Scientific Questions. In *ICLR 2025 Workshop on Navigating and Addressing Data Problems for Foundation Models*.
[8] Tiejin Chen, Xiaoou Liu, Longchao Da, Jia Chen, Vagelis Papalexakis, and Hua Wei. 2025. Uncertainty Quantification of Large Language Models through Multi-Dimensional Responses. *arXiv preprint arXiv:2502.16820* (2025).
[9] Zizhang Chen, Peizhao Li, Xiaomeng Dong, and Pengyu Hong. 2024. Uncertainty Quantification for Clinical Outcome Predictions with (Large) Language Models. *arXiv preprint arXiv:2411.03497* (2024).

[10] Inyoung Cheong, King Xia, KJ Kevin Feng, Quan Ze Chen, and Amy X Zhang. 2024. (A) I am not A lawyer, but...: engaging legal experts towards responsible LLM policies for legal advice. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. 2454–2469.

[11] John Cherian, Isaac Gibbs, and Emmanuel Candes. 2025. Large language model validity via enhanced conformal prediction methods. *Advances in Neural Information Processing Systems* 37 (2025), 114812–114842.

[12] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training Verifiers to Solve Math Word Problems. *arXiv preprint arXiv:2110.14168* (2021).

[13] Jeremy Cole, Michael Zhang, Dan Gillick, Julian Eisenschlos, Bhuwan Dhingra, and Jacob Eisenstein. 2023. Selectively Answering Ambiguous Questions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 530–543.

[14] Antonia Creswell, Murray Shanahan, and Irina Higgins. [n. d.]. Selection-Inference: Exploiting Large Language Models for Interpretable Logical Reasoning. In *The Eleventh International Conference on Learning Representations*.

[15] Longchao Da, Tiejin Chen, Lu Cheng, and Hua Wei. 2024. Llm uncertainty quantification through directional entailment graph and claim level response augmentation. *arXiv preprint arXiv:2407.00994* (2024).

[16] Longchao Da, Minquan Gao, Hao Mei, and Hua Wei. 2024. Prompt to transfer: Sim-to-real transfer for traffic signal control with prompt learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 82–90.

[17] Longchao Da, Kuanru Liou, Tiejin Chen, Xuesong Zhou, Xiangyong Luo, Yezhou Yang, and Hua Wei. 2024. Open-ti: Open traffic intelligence with augmented language model. *International Journal of Machine Learning and Cybernetics* 15, 10 (2024), 4761–4786.

[18] Longchao Da, Xiaoou Liu, Jiaxin Dai, Lu Cheng, Yaqing Wang, and Hua Wei. 2025. Understanding the Uncertainty of LLM Explanations: A Perspective Based on Reasoning Topology. *arXiv preprint arXiv:2502.17026* (2025).

[19] Longchao Da, Hao Mei, Romir Sharma, and Hua Wei. 2023. Uncertainty-aware grounded action transformation towards sim-to-real transfer for traffic signal control. In *2023 62nd IEEE Conference on Decision and Control (CDC)*. 1124–1129.

[20] Longchao Da, Rui Wang, Xiaojian Xu, Parminder Bhatia, Taha Kass-Hout, Hua Wei, and Cao Xiao. 2024. Segment as You Wish–Free-Form Language-Based Segmentation for Medical Images. *arXiv preprint arXiv:2410.12831* (2024).

[21] Irene de Zarzà, Joachim de Curtò, Gemma Roig, and Carlos T Calafate. 2023. LLM multimodal traffic accident forecasting. *Sensors* 23, 22 (2023), 9225.

[22] Yang Deng, Shuaiyi Li, and Wai Lam. 2023. Learning to ask clarification questions with spatial reasoning. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2113–2117.

[23] Armen Der Kiureghian and Ove Ditlevsen. 2009. Aleatory or epistemic? Does it matter? *Structural safety* 31, 2 (2009), 105–112.

[24] Jinhao Duan, Hao Cheng, Shiqi Wang, Alex Zavalny, Chenan Wang, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. 2024. Shifting Attention to Relevance: Towards the Predictive Uncertainty Quantification of Free-Form Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 5050–5063.

[25] Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. 2021. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics* 9 (2021), 1012–1031.

[26] Ekaterina Fadeeva, Aleksandr Rubashevskii, Artem Shelmanov, Sergey Petrakov, Haonan Li, Hamdy Mubarak, Evgenii Tsymbalov, Gleb Kuzmin, Alexander Panchenko, Timothy Baldwin, et al. 2024. Fact-Checking the Output of Large Language Models via Token-Level Uncertainty Quantification. In *Findings of the Association for Computational Linguistics ACL 2024*. 9367–9385.

[27] Ekaterina Fadeeva, Roman Vashurin, Akim Tsvigun, Artem Vazhentsev, Sergey Petrakov, Kirill Fedyanin, Daniil Vasilev, Elizaveta Goncharova, Alexander Panchenko, Maxim Panov, et al. 2023. LM-Polygraph: Uncertainty Estimation for Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 446–461.

[28] Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*. PMLR, 1050–1059.

[29] Xiang Gao, Jiaxin Zhang, Lalla Mouatadid, and Kamalika Das. 2024. SPUQ: Perturbation-Based Uncertainty Quantification for Large Language Models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2336–2346.

[30] Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did Aristotle Use a Laptop? A Question Answering Benchmark with Implicit Reasoning Strategies. *Transactions of the Association for Computational Linguistics* (2021), 346–361.

[31] Mario Giulianelli, Joris Baan, Wilker Aziz, Raquel Fernández, and Barbara Plank. 2023. What Comes Next? Evaluating Uncertainty in Neural Text Generators Against Human Production Variability. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 14349–14371.

[32] Olga Golovneva, Moya Peng Chen, Spencer Poff, Martin Corredor, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. [n. d.]. ROSCOE: A Suite of Metrics for Scoring Step-by-Step Reasoning. In *The Eleventh International Conference on Learning Representations*.

[33] Meiqi Guo, Mingda Zhang, Siva Reddy, and Malihe Alikhani. 2021. Abg-coqa: Clarifying ambiguity in conversational question answering. *3rd Conference on Automated Knowledge Base Construction* (2021).

[34] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300* (2020).

[35] Jay Heo, Hae Beom Lee, Saehoon Kim, Juho Lee, Kwang Joon Kim, Eunho Yang, and Sung Ju Hwang. 2018. Uncertainty-aware attention for reliable interpretation and prediction. *Advances in neural information processing systems* 31 (2018).

[36] Bairu Hou, Yujian Liu, Kaizhi Qian, Jacob Andreas, Shiyu Chang, and Yang Zhang. 2024. Decomposing Uncertainty for Large Language Models through Input Clarification Ensembling. In *Forty-first International Conference on Machine Learning*.

[37] Hsiu-Yuan Huang, Yutong Yang, Zhaoxi Zhang, Sanwoo Lee, and Yunfang Wu. 2024. A survey of uncertainty estimation in llms: Theory meets practice. *arXiv preprint arXiv:2410.15326* (2024).

[38] Jingwang Huang, Jiang Zhong, Qin Lei, Jinpeng Gao, Yuming Yang, Sirui Wang, Peiguang Li, and Kaiwen Wei. 2025. Latent Distribution Decoupling: A Probabilistic Framework for Uncertainty-Aware Multimodal Emotion Recognition. *arXiv preprint arXiv:2502.13954* (2025).

[39] Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. *arXiv preprint arXiv:1909.00277* (2019).

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

[41] Yukun Huang, Yixin Liu, Raghuveer Thirukovalluru, Arman Cohan, and Bhuwan Dhingra. 2024. Calibrating Long-form Generations From Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. 13441–13460.

[42] Eyke Hüllermeier and Willem Waegeman. 2021. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine learning* 110, 3 (2021), 457–506.

[43] Fushuo Huo, Wenchao Xu, Zhong Zhang, Haozhao Wang, Zhicheng Chen, and Peilin Zhao. 2024. Self-introspective decoding: Alleviating hallucinations for large vision-language models. *arXiv preprint arXiv:2408.02032* (2024).

[44] Abhyuday Jagannatha and Hong Yu. 2020. Calibrating Structured Output Predictors for Natural Language Processing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2078–2092.

[45] Heinrich Jiang, Been Kim, Maya Gupta, and Melody Y. Guan. 2018. To trust or not to trust a classifier. In *Advances in Neural Information Processing Systems*.

[46] Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021. How Can We Know When Language Models Know? On the Calibration of Language Models for Question Answering. *Transactions of the Association for Computational Linguistics* (2021), 962–977.

[47] Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences* 11, 14 (2021), 6421.

[48] Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*. 1601–1611.

[49] 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* (2022).

[50] Sanyam Kapoor, Nate Gruver, Manley Roberts, Arka Pal, Samuel Dooley, Micah Goldblum, and Andrew Wilson. 2024. Calibration-Tuning: Teaching Large Language Models to Know What They Don't Know. In *Proceedings of the 1st Workshop on Uncertainty-Aware NLP (UncertaiNLP 2024)*. 1–14.

[51] Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation. In *The Eleventh International Conference on Learning Representations*.

[52] Aviral Kumar and Sunita Sarawagi. 2019. Calibration of encoder decoder models for neural machine translation. *arXiv preprint arXiv:1903.00802* (2019).

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

[54] Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding Comprehension Dataset From Examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 785–794.

[55] Siqi Lai, Zhao Xu, Weijia Zhang, Hao Liu, and Hui Xiong. 2025. LLMLight: Large Language Models as Traffic Signal Control Agents. *31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (2025).

[56] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems* 30 (2017).

[57] Baolin Li, Yankai Jiang, Vijay Gadepally, and Devesh Tiwari. 2024. Llm inference serving: Survey of recent advances and opportunities. *arXiv preprint arXiv:2407.12391* (2024).

[58] Junyi Li, Xiaoxue Cheng, Wayne 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*. 6449–6464.

[59] Lincan Li, Jiaqi Li, Catherine Chen, Fred Gui, Hongjia Yang, Chenxiao Yu, Zhengguang Wang, Jianing Cai, Junlong Aaron Zhou, Bolin Shen, et al. 2024. Political-llm: Large language models in political science. *arXiv preprint arXiv:2412.06864* (2024).

[60] Yaowei Li, Bang Yang, Xuxin Cheng, Zhihong Zhu, Hongxiang Li, and Yuexian Zou. 2023. Unify, align and refine: Multi-level semantic alignment for radiology report generation. In *Proceedings of the IEEE/CVF international conference on computer vision*. 2863–2874.

[61] Zixuan Li, Jing Xiong, Fanghua Ye, Chuanyang Zheng, Xun Wu, Jianqiao Lu, Zhongwei Wan, Xiaodan Liang, Chengming Li, Zhenan Sun, et al. 2024. UncertaintyRAG: Span-Level Uncertainty Enhanced Long-Context Modeling for Retrieval-Augmented Generation. *arXiv preprint arXiv:2410.02719* (2024).

[62] Kaiqu Liang, Zixu Zhang, and Jaime Fisac. 2024. Introspective Planning: Aligning Robots' Uncertainty with Inherent Task Ambiguity. *Advances in Neural Information Processing Systems* 37 (2024), 71998–72031.

[63] Siyuan Liang, Jiawei Liang, Tianyu Pang, Chao Du, Aishan Liu, Ee-Chien Chang, and Xiaochun Cao. 2024. Revisiting backdoor attacks against large vision-language models. *arXiv preprint arXiv:2406.18844* (2024).

[64] Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching models to express their uncertainty in words. *arXiv preprint arXiv:2205.14334* (2022).

[65] Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring How Models Mimic Human Falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 3214–3252.

[66] Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2023. Generating with Confidence: Uncertainty Quantification for Black-box Large Language Models. *Transactions on Machine Learning Research* (2023).

[67] Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2024. Contextualized Sequence Likelihood: Enhanced Confidence Scores for Natural Language Generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 10351–10368.

[68] Chen Ling, Xujiang Zhao, Xuchao Zhang, Wei Cheng, Yanchi Liu, Yiyou Sun, Mika Oishi, Takao Osaki, Katsushi Matsuda, Jie Ji, Guangji Bai, Liang Zhao, and Haifeng Chen. 2024. Uncertainty Quantification for In-Context Learning of 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)*. 3357–3370.

[69] Linyu Liu, Yu Pan, Xiaocheng Li, and Guanting Chen. 2024. Uncertainty estimation and quantification for llms: A simple supervised approach. *arXiv preprint arXiv:2404.15993* (2024).

[70] Shudong Liu, Zhaocong Li, Xuebo Liu, Runzhe Zhan, Derek Wong, Lidia Chao, and Min Zhang. 2024. Can LLMs learn uncertainty on their own? expressing uncertainty effectively in a self-training manner. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 21635–21645.

[71] Xin Liu, Muhammad Khalifa, and Lu Wang. 2024. LitCab: Lightweight Language Model Calibration over Short- and Long-form Responses. In *The Twelfth International Conference on Learning Representations*.

[72] Xiaoou Liu, Zhen Lin, Longchao Da, Chacha Chen, Shubhendu Trivedi, and Hua Wei. 2025. MCQA-Eval: Efficient Confidence Evaluation in NLG with Gold-Standard Correctness Labels. *arXiv preprint arXiv:2502.14268* (2025).

[73] Jinliang Lu, Ziliang Pang, Min Xiao, Yaochen Zhu, Rui Xia, and Jiajun Zhang. 2024. Merge, ensemble, and cooperate! a survey on collaborative strategies in the era of large language models. *arXiv preprint arXiv:2407.06089* (2024).

[74] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2023. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems* 36 (2023), 46534–46594.

[75] Andrey Malinin and Mark Gales. 2021. Uncertainty Estimation in Autoregressive Structured Prediction. In *International Conference on Learning Representations*.

[76] Potsawee Manakul, Adian Liusie, and Mark Gales. [n. d.]. SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models. In *2023 Conference on Empirical Methods in Natural Language Processing*.

[77] Katerina Margatina, Timo Schick, Nikolaos Aletras, and Jane Dwivedi-Yu. 2023. Active Learning Principles for In-Context Learning with Large Language Models. In *Findings of the Association for Computational Linguistics*. 5011–5034.

[78] Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit yyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 12076–12100.

[79] Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. AmbigQA: Answering Ambiguous Open-domain Questions. In *EMNLP*.

[80] Shentong Mo and Miao Xin. 2024. Tree of uncertain thoughts reasoning for large language models. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 12742–12746.

[81] Philipp Mondorf and Barbara Plank. 2024. Beyond Accuracy: Evaluating the Reasoning Behavior of Large Language Models-A Survey. In *First Conference on Language Modeling*.

[82] 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)*. 368–378.

[83] Subhabrata Mukherjee and Ahmed Awadallah. 2020. Uncertainty-aware self-training for few-shot text classification. *Advances in Neural Information Processing Systems* 33 (2020), 21199–21212.

[84] James F Mullen Jr and Dinesh Manocha. 2024. LAP, Using Action Feasibility for Improved Uncertainty Alignment of Large Language Model Planners. *arXiv preprint arXiv:2403.13198* (2024).

[85] Allan H. Murphy and Robert L. Winkler. 1977. Reliability of Subjective Probability Forecasts of Precipitation and Temperature. *Journal of The Royal Statistical Society Series C-applied Statistics* 26 (1977), 41–47.

[86] Malik Sajjad Ahmed Nadeem, Jean-Daniel Zucker, and Blaise Hanczar. 2009. Accuracy-rejection curves (ARCs) for comparing classification methods with a reject option. In *Machine Learning in Systems Biology*. 65–81.

[87] Alexander Nikitin, Jannik Kossen, Yarin Gal, and Pekka Marttinen. 2024. Kernel language entropy: Fine-grained uncertainty quantification for LLMs from semantic similarities. *Advances in Neural Information Processing Systems* 37 (2024), 8901–8929.

[88] Hyobin Ong, Youngwoo Yoon, Jaewoo Choi, and Minsu Jang. 2024. A Simple Baseline for Uncertainty-Aware Language-Oriented Task Planner for Embodied Agents. In *2024 21st International Conference on Ubiquitous Robots (UR)*. 677–682.

[89] Arjun Panickssery, Samuel Bowman, and Shi Feng. 2024. Llm evaluators recognize and favor their own generations. *Advances in Neural Information Processing Systems* 37 (2024), 68772–68802.

[90] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 311–318.

[91] Jianing Qiu, Kyle Lam, Guohao Li, Amish Acharya, Tien Yin Wong, Ara Darzi, Wu Yuan, and Eric J Topol. 2024. LLM-based agentic systems in medicine and healthcare. *Nature Machine Intelligence* 6, 12 (2024), 1418–1420.

[92] Victor Quach, Adam Fisch, Tal Schuster, Adam Yala, Jae Ho Sohn, Tommi S Jaakkola, and Regina Barzilay. [n. d.]. Conformal Language Modeling. In *The Twelfth International Conference on Learning Representations*.

[93] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250* (2016).

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

[95] David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. 2024. Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.

[96] 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. In *The Eleventh International Conference on Learning Representations*.

[97] Jie Ren, Yao Zhao, Tu Vu, Peter J. Liu, and Balaji Lakshminarayanan. 2023. Self-Evaluation Improves Selective Generation in Large Language Models. In *Proceedings on "I Can't Believe It's Not Better: Failure Modes in the Age of Foundation Models" at NeurIPS 2023 Workshops*, Vol. 239. 49–64.

[98] Thomas Savage, John Wang, Robert Gallo, Abdessalem Boukil, Vishwesh Patel, Seyed Amir Ahmad Safavi-Naini, Ali Soroush, and Jonathan H Chen. 2024. Large language model uncertainty measurement and calibration for medical diagnosis and treatment. *medRxiv* (2024), 2024–06.

[99] Sagar Sen, Victor Gonzalez, Erik Johannes Husom, Simeon Tverdal, Shukun Tokas, and Svein O Tjøsvoll. 2024. ERG-AI: enhancing occupational ergonomics with uncertainty-aware ML and LLM feedback. *Applied Intelligence* 54, 23 (2024),

12128–12155.

[100] Murat Sensoy, Lance Kaplan, and Melih Kandemir. 2018. Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems* 31 (2018).

[101] Glenn Shafer and Vladimir Vovk. 2008. A tutorial on conformal prediction. *Journal of Machine Learning Research* 9, 3 (2008).

[102] Ola Shorinwa, Zhiting Mei, Justin Lidard, Allen Z Ren, and Anirudha Majumdar. 2024. A survey on uncertainty quantification of large language models: Taxonomy, open research challenges, and future directions. *arXiv preprint arXiv:2412.05563* (2024).

[103] Elias Stengel-Eskin and Benjamin Van Durme. 2023. Calibrated Interpretation: Confidence Estimation in Semantic Parsing. *Transactions of the Association for Computational Linguistics* 11 (2023), 1213–1231.

[104] Jiayuan Su, Jing Luo, Hongwei Wang, and Lu Cheng. 2024. API Is Enough: Conformal Prediction for Large Language Models Without Logit-Access. In *Findings of the Association for Computational Linguistics: EMNLP 2024.* 979–995.

[105] Wei Suo, Lijun Zhang, Mengyang Sun, Lin Yuanbo Wu, Peng Wang, and Yanning Zhang. 2025. Octopus: Alleviating Hallucination via Dynamic Contrastive Decoding. *arXiv preprint arXiv:2503.00361* (2025).

[106] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. *arXiv preprint arXiv:1803.05355* (2018).

[107] Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. 2023. Just Ask for Calibration: Strategies for Eliciting Calibrated Confidence Scores from Language Models Fine-Tuned with Human Feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing.* 5433–5442.

[108] Yao-Hung Hubert Tsai, Walter Talbott, and Jian Zhang. 2024. Efficient Non-Parametric Uncertainty Quantification for Black-Box Large Language Models and Decision Planning. *arXiv preprint arXiv:2402.00251* (2024).

[109] Dennis Ulmer, Martin Gubri, Hwaran Lee, Sangdoo Yun, and Seong Oh. 2024. Calibrating Large Language Models Using Their Generations Only. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).* 15440–15459.

[110] Roman Vashurin, Ekaterina Fadeeva, Artem Vazhentsev, Lyudmila Rvanova, Daniil Vasilev, Akim Tsvigun, Sergey Petrakov, Rui Xing, Abdelrahman Sadallah, Kirill Grishchenkov, Alexander Panchenko, Timothy Baldwin, Preslav Nakov, Maxim Panov, and Artem Shelmanov. 2025. Benchmarking Uncertainty Quantification Methods for Large Language Models with LM-Polygraph. *Transactions of the Association for Computational Linguistics* (2025), 220–248.

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

[112] 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).* 11659–11681.

[113] Boshi Wang, Xiang Yue, Yu Su, and Huan Sun. 2024. Grokking of implicit reasoning in transformers: A mechanistic journey to the edge of generalization. *Advances in Neural Information Processing Systems* 37 (2024), 95238–95265.

[114] Shuo Wang, Zhaopeng Tu, Shuming Shi, and Yang Liu. 2020. On the Inference Calibration of Neural Machine Translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.*

[115] Yibin Wang, Haizhou Shi, Ligong Han, Dimitris Metaxas, and Hao Wang. 2025. Blob: Bayesian low-rank adaptation by backpropagation for large language models. *Advances in Neural Information Processing Systems* 37 (2025), 67758–67794.

[116] Zhiyuan Wang, Jinhao Duan, Lu Cheng, Yue Zhang, Qingni Wang, Xiaoshuang Shi, Kaidi Xu, Heng Tao Shen, and Xiaofeng Zhu. 2024. ConU: Conformal Uncertainty in Large Language Models with Correctness Coverage Guarantees. In *Findings of the Association for Computational Linguistics.* 6886–6898.

[117] Jiaxin Wu, Yizhou Yu, and Hong-Yu Zhou. 2024. Uncertainty Estimation of Large Language Models in Medical Question Answering. *arXiv preprint arXiv:2407.08662* (2024).

[118] Shuo Xing, Yuping Wang, Peiran Li, Ruizheng Bai, Yueqi Wang, Chan-wei Hu, Chengxuan Qian, Huaxiu Yao, and Zhengzhong Tu. 2025. Re-Align: Aligning Vision Language Models via Retrieval-Augmented Direct Preference Optimization. *arXiv preprint arXiv:2502.13146* (2025).

[119] 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. In *The Twelfth International Conference on Learning Representations.*

[120] Miao Xiong, Andrea Santilli, Michael Kirchhof, Adam Golinski, and Sinead Williamson. 2024. Efficient and effective uncertainty quantification for LLMs. In *Neurips Safe Generative AI Workshop 2024.*

[121] Adam X Yang, Maxime Robeyns, Xi Wang, and Laurence Aitchison. [n. d.]. Bayesian Low-rank Adaptation for Large Language Models. In *The Twelfth*

[122] Yongjin Yang, Haneul Yoo, and Hwaran Lee. 2025. MAQA: Evaluating Uncertainty Quantification in LLMs Regarding Data Uncertainty. In *Findings of the Association for Computational Linguistics: NAACL 2025.* 5846–5863.

[123] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600* (2018).

[124] Huaiyuan Yao, Longchao Da, Vishnu Nandam, Justin Turnau, Zhiwei Liu, Linsey Pang, and Hua Wei. 2025. Comal: Collaborative multi-agent large language models for mixed-autonomy traffic. In *Proceedings of the 2025 SIAM International Conference on Data Mining (SDM).* 409–418.

[125] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems* 36 (2023), 11809–11822.

[126] Fanghua Ye, Mingming Yang, Jianhui Pang, Longyue Wang, Derek Wong, Emine Yilmaz, Shuming Shi, and Zhaopeng Tu. 2025. Benchmarking llms via uncertainty quantification. *Advances in Neural Information Processing Systems* (2025), 15356–15385.

[127] Kai Ye, Tiejin Chen, Hua Wei, and Liang Zhan. 2024. Uncertainty regularized evidential regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 16460–16468.

[128] Zhangyue Yin, Qiushi Sun, Qipeng Guo, Zhiyuan Zeng, Xiaonan Li, Junqi Dai, Qinyuan Cheng, Xuan-Jing Huang, and Xipeng Qiu. 2024. Reasoning in flux: Enhancing large language models reasoning through uncertainty-aware adaptive guidance. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).* 2401–2416.

[129] Spencer Young, Porter Jenkins, Lonchao Da, Jeff Dotson, and Hua Wei. 2025. Flexible heteroscedastic count regression with deep double poisson networks. *International Conference on Machine Learning* (2025).

[130] Hamed Zamani, Susan Dumais, Nick Craswell, Paul Bennett, and Gord Lueck. 2020. Generating clarifying questions for information retrieval. In *Proceedings of the web conference 2020.* 418–428.

[131] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a Machine Really Finish Your Sentence?. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics.* 4791–4800.

[132] Boxuan Zhang and Ruqi Zhang. 2025. CoT-UQ: Improving Response-wise Uncertainty Quantification in LLMs with Chain-of-Thought. *arXiv preprint arXiv:2502.17214* (2025).

[133] Caiqi Zhang, Fangyu Liu, Marco Basaldella, and Nigel Collier. 2024. LUQ: Long-text Uncertainty Quantification for LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing.* 5244–5262.

[134] Ruiyang Zhang, Hu Zhang, and Zhedong Zheng. 2024. VL-Uncertainty: Detecting Hallucination in Large Vision-Language Model via Uncertainty Estimation. *arXiv preprint arXiv:2411.11919* (2024).

[135] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. [n. d.]. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations.*

[136] Yuan Zhang, Tao Huang, Chun-Kai Fan, Hongyuan Dong, Jiawen Li, Jiacong Wang, Kuan Cheng, Shanghang Zhang, Haoyuan Guo, et al. 2024. Unveiling the tapestry of consistency in large vision-language models. *Advances in Neural Information Processing Systems* 37 (2024), 118632–118653.

[137] Zheng Zhang, Xu Yuan, Lei Zhu, Jingkuan Song, and Liqiang Nie. 2024. BadCM: Invisible backdoor attack against cross-modal learning. *IEEE Transactions on Image Processing* (2024).

[138] Guosheng Zhao, Zijian Zhao, Wuxian Gong, and Feng Li. 2023. Radiology report generation with medical knowledge and multilevel image-report alignment: A new method and its verification. *Artificial Intelligence in Medicine* 146 (2023), 102714.

[139] Yao Zhao, Mikhail Khalman, Rishabh Joshi, Shashi Narayan, Mohammad Saleh, and Peter J Liu. 2023. Calibrating Sequence likelihood Improves Conditional Language Generation. In *The Eleventh International Conference on Learning Representations.*

[140] Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2023. Large language models are not robust multiple choice selectors. *arXiv preprint arXiv:2309.03882* (2023).

[141] Zhi Zheng, Qian Feng, Hang Li, Alois Knoll, and Jianxiang Feng. 2024. Evaluating uncertainty-based failure detection for closed-loop llm planners. *arXiv preprint arXiv:2406.00430* (2024).