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

Xiaoou Liu*
Arizona State University
xiaoouli@asu.edu

Tiejin Chen*
Arizona State University
tchen169@asu.edu

Longchao Da
Arizona State University
longchao@asu.edu

Chacha Chen
University of Chicago
chacha@uchicago.edu

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

Hua Wei
Arizona State University
hua.wei@asu.edu

Abstract

Large Language Models (LLMs) excel in text generation, reasoning, and decision-making, enabling their adoption in high-stakes domains such as healthcare, law, and transportation. However, their reliability is a major concern, as they often produce plausible but incorrect responses. Uncertainty quantification (UQ) enhances trustworthiness 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 (input, reasoning, parameter, and prediction uncertainty). We evaluate existing techniques, assess their real-world applicability, and identify open challenges, emphasizing the need for scalable, interpretable, and robust UQ approaches to enhance LLM reliability.

Keywords

Uncertainty Quantification; Large Language Models

ACM Reference Format:
Xiaoou Liu*, Tiejin Chen*, Longchao Da, Chacha Chen, Zhen Lin, and Hua Wei. 2025. Uncertainty Quantification and Confidence Calibration in Large Language Models: A Survey. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2 (KDD '25), August 3–7, 2025, Toronto, ON, Canada.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3711896.3736569

1 Introduction

Large Language Models (LLMs) like GPT-4 [1] and PaLM [2] have achieved remarkable capabilities in text generation, reasoning, and decision-making, driving their adoption in high-stakes domains such as healthcare diagnostics [24, 38, 105, 128], legal analysis [12, 29, 54], and transportation systems [18, 58, 67, 165]. 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 [57]. 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 a pivotal mechanism to enhance LLM trustworthiness 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 [83, 120]. For instance, in clinical settings, uncertainty-aware LLMs could flag uncertain diagnoses for specialist review, reducing diagnostic errors by up to 41% [116]. 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 Large Language Models (LLMs). Bayesian approaches like Monte Carlo dropout [34] are computationally prohibitive for trillion-parameter models and natural language generation (NLG) tasks, while ensemble methods struggle with consistency across diverse decoding strategies [85]. Furthermore, LLMs introduce unique uncertainty sources—such as input ambiguity [9, 41], reasoning path divergence, and decoding stochasticity—that transcend classical aleatoric and epistemic categorizations [51]. 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.

This tutorial introduces a novel taxonomy for LLM uncertainty quantification, categorizing methods along two axes: (1) computational efficiency (e.g., single-pass vs. sampling-based techniques) and (2) uncertainty dimensions (input, reasoning, parametric, predictive). Our framework addresses three gaps in prior work: 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,

KDD '25, Toronto, ON, Canada

ACM ISBN 979-8-4007-1454-2/2025/08
https://doi.org/10.1145/3711896.3736569

epistemic uncertainty, or a mixture of both. Third, it identifies understudied areas like reasoning uncertainty, which accounts for 58% of errors in multi-step QA tasks [20].

Contributions: We provide (1) the first systematic review of UQ methods tailored to different dimensions of uncertainty in LLM, including input, reasoning, parameter, and prediction uncertainty; (2) a comprehensive introduction to the evaluation of UQ methods and the challenges.

Connection to Existing Surveys: Prior surveys [7, 45, 48, 108, 118, 157] focus narrowly 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—factors 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. Sections 5 and 6 introduce the applications of UQ in different domains with LLMs and identify open challenges and future directions. Table 1 shows the notation and Table 2 shows an overview of all methods we discuss in this paper.

2 Perliminaries

2.1 Sources of Uncertainty in LLMs

2.1.1 Aleatoric vs. Epistemic Uncertainty. For uncertainty quantification of traditional machine learning tasks such as classification or regression [149], there are mainly two types of uncertainty [28]: aleatoric and epistemic uncertainty. Aleatoric uncertainty defines the uncertainty from noise in the dataset. Epistemic uncertainty, on the other hand, 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. Furthermore, ambiguity in natural language 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.

Though uncertainty for LLMs can also be classified as these two categories, these two categories alone are insufficient to fully capture the complexities of uncertainty in the language model. The unique nature of LLM inference introduces additional uncertainty factors. In particular, LLMs exhibit uncertainty not only due to training data limitations but also due to input variability and decoding mechanisms. Therefore, to address these challenges, we introduce four new dimensions of uncertainty.

2.1.2 Uncertainty with Different Dimensions. To systematically analyze uncertainty in LLMs, we categorize it into four key dimensions: input uncertainty, reasoning uncertainty, parameter uncertainty, and prediction uncertainty. Each dimension may involve aleatoric uncertainty, epistemic uncertainty, or a combination of both. This structured framework provides a more comprehensive understanding of uncertainty quantification in LLMs.

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

Table 1: Notation used in this paper.

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 logic or retrieval, leading to ambiguous or incorrect reasoning. 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 gaps in the training data, 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 [34], deep ensembles [68], and uncertainty-aware training [96] 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

Uncertainty quantification and confidence estimation are closely related yet distinct concepts in the context of large language models (LLMs). 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 answer or prediction. As a concrete example, in the context of classification, a simple confidence measure is the predicted probability $\hat{p}(Y = y|x)$ (an uncertainty measure which does not depend on the particular prediction y could be entropy, taking the form of $\sum_y -\hat{p}(Y = y|x) \log \hat{p}(Y = y|x)$). The corresponding **confidence score** in NLG (for an auto-regressive LM) is the joint probability for the generated sequence:

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

The log of Eq. (1) is sometimes referred to as *sequence likelihood* [162]. In general, while an uncertainty estimate takes the form of $U(x)$, confidence estimates could be expressed as $C(\mathbf{x}, \mathbf{s})$. Note that unlike classification tasks, not 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 $f(\mathbf{s}|\mathbf{x})$.

There are usually two dimensions along which researchers improve confidence estimates in NLG, which is unsurprisingly largely influenced by confidence scoring literature from classification [16, 55], 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 [60] or AUARC [79] 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 LLM or even modern machine learning [26, 98, 99], but bears slightly different meanings in NLP. In general, we could define a perfectly calibrated confidence measure to achieve the following:

$$\forall c, \mathbb{E}[f(\mathbf{s}|\mathbf{x})|C(\mathbf{x}, \mathbf{s}) = c] = c, \quad (2)$$

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 LMs and tasks [62, 64, 135]. Evaluation typically relies on variants of Expected Calibration Error (ECE) [64, 72, 125]. Oftentimes confidence scores from classification could be directly applied [56, 121, 163] 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.

As uncertainty and confidence are often intertwined, many approaches used in uncertainty quantification have their counterpart in confidence estimation. For example, for black-box methods, Lin et al. [78] 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. Zhang et al. [154] extends such black-box methods to longer generations. For logit-based methods, Malinin and Gales [87] normalize Eq. (1) with the length of $\mathbf{s}$. Further improvements include replacing the logit-sum or mean with weighted sum, by attention values on the important tokens [79] or by importance inferred from natural language inference (NLI) models [30]. Such variants of sequence likelihood could then be fed for (entropy-style) uncertainty computation [63, 79].

Another popular approach is asking the LM itself whether a particular free-form generation is correct [60]. 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 Ren et al. [113], which generalizes Kadavath et al. [60] and converts samples from free-form generation into a *multiple-choice* question (with generations being the options) and adding a "None of the above" option to elicit the confidence. Similarly, Shrivastava et al. [119] performs the explicit relative comparison of pairs of answers and then aggregates the preferences into an absolute confidence level.

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 full LM [56, 61, 162] or partial [83]) with a different loss, or using a separate model on the internal or external representations from the base LM [3, 53, 90, 112, 129]. On the other end of the spectrum, without any training, prompting could be used to elicit verbalized confidence values [125, 140], or to recalibrate LLM confidence for a particular distribution [72] via in-context learning. Finally, one could combine multiple confidence estimation methods and enjoy the benefit of ensembling [35].

Just like the evaluation for uncertainty quantification (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 [78, 83, 125], human annotations [90, 113], or lexical similarities such as ROUGE [63, 162]. Recently, Liu et al. [84] proposes to evaluate free-form generation confidence measures with selected multiple-choice datasets as an efficient complement. For longer generations, Huang et al. [50] proposes to use ordinal (not binary) correctness values to capture the ambiguity in the quality of a generation. In a similar flavor, Baan et al. [4] studies the issues in the evaluation of calibration (against the human majority) 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 [106] or *seletive generation* [15]). In the rest of this paper, we sometimes follow the language of the original papers and treat confidence

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

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 and many works in the LLMs domain try to deal with ambiguity. For example, there are many datasets on ambiguity with different tasks such as ambiguity detection [41, 92] and many works on dealing with ambiguity [27, 150]. However, these works did not consider uncertainty at all.

As far as we know, most uncertainty quantification methods that care about input uncertainty focus on perturbing the input prompts of LLMs. For instance, Hou et al. [44] 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, Ling et al. [80] quantified the input uncertainty in the setting of in-context learning by using different in-context samples. Gao et al. [35] 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 [156] to measure how consistent the responses are across different perturbations.

We can see that there are only a few papers that care about input uncertainty and its application to ambiguity. Since ambiguity is common and important in natural language, we call more effort into input uncertainty and its application.

3.2 Reasoning Uncertainty

Recent research has advanced stepwise uncertainty quantification in LLM reasoning by explicitly eliciting and analyzing the internal reasoning process. For example, Tree of Uncertain Thoughts (TouT) [94] extends the Tree of Thoughts (ToT) [146] framework by addressing the inherent local uncertainties that occur during intermediate reasoning steps. TouT leverages Monte Carlo Dropout to assign uncertainty scores to pivotal decision points and, by integrating these local measures with global search techniques, it enhances the precision of response generation. Similarly, TopologyUQ [20] 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. In addition, Uncertainty-aware Adaptive Guidance (UAG) [148] tackles 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. Complementing these methods, Stable-Explanation Confidence [6] quantifies uncertainty by examining the distribution of generated explanations, treating each model's explanation pair as a test-time classifier to construct a posterior answer distribution that reflects overall reasoning confidence. More recently, CoT-UQ [153] has integrated chain-of-thought reasoning into a response-level uncertainty quantification 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 both local and global levels.

3.3 Parameter Uncertainty

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

Traditional UQ methods like Monte Carlo (MC) 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)[137] and Bayesian Low-Rank Adaptation (BLoRA)[142] incorporate Bayesian modeling into LoRA adapters, allowing uncertainty estimation through parameter distributions without 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[81] train auxiliary models to predict the confidence of LLM outputs based on activation patterns and logit distributions. Similarly, Uncertainty-aware Instruction Tuning (UaIT)[82] modifies the fine-tuning process to explicitly train models to express uncertainty in their outputs. SAPLMA[3] refines probabilistic alignment techniques to dynamically adjust model uncertainty estimates, ensuring adaptability to different downstream tasks. Additionally, LoRA ensembles[5] 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 uncertainty quantification methods focus on prediction uncertainty since it is the most straightforward way to estimate the uncertainty.

3.4.1 Single Round Generation. Most single-round generation methods utilize the logit or hidden states during the generation procedure. Single-round generation methods have the highest efficiency.

Perplexity is a measure of how well a probabilistic language model predicts a sequence of text [131] while Mora-Cross and Calderon-Ramirez [95], Margatina et al. [89] and Manakul et al. [88] utilize the perplexity as the uncertainty. In detail, using w_i as the i-th token in the generation, perplexity is given by:

$$\text{Perplexity} = \exp\left(-\frac{1}{N}\sum_{i=1}^{N}\ln p(w_i)\right) \tag{3}$$

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, Manakul et al. [88] also uses maximum token log-probability:

$$Max(p) = \max_i(-\ln p(w_i)) \tag{4}$$

Method	Uncertainty Stages	Efficency Features	Access to model	Confidence
Input clarification ensembles [44]	Input Uncertainty	Multi Rounds Generations	Black-box	No
Uncertainty Quantification for In-Context Learning [80]	Input Uncertainty	Multi Rounds Generations	Black-box	No
SPUQ [35]	Input Uncertainty	Multi Rounds Generations + Additional Model	Black-box	No
UAG [148]	Reasoning Uncertainty	Single Round Generation	White-box	No
CoT-UQ [153]	Reasoning Uncertainty	Single Round Generation	White-box	Yes
TouT [94]	Reasoning Uncertainty	Multi Rounds Generations	Black-box	No
TopologyUQ [20]	Reasoning Uncertainty	Multi Rounds Generations	Black-box	No
Stable Explanations Confidence [6]	Reasoning Uncertainty	Multi Rounds Generation	Black-box	Yes
SAPLMA [3]	Parameter + Prediction Uncertainty	Fine-tuning	White-box	Yes
Supervised uncertainty estimation[81]	Parameter + Prediction Uncertainty	Fine-tuning	White-box	Yes
UaIT [82]	Parameter + Prediction Uncertainty	Fine-tuning	White-box	Yes
LoRA ensembles [5]	Parameter Uncertainty	Fine-tuning	White-box	Yes
BloB [137]	Parameter Uncertainty	Fine-tuning	White-box	Yes
BLoRA [142]	Parameter Uncertainty	Fine-tuning	White-box	Yes
Perplexity [89, 95]	Prediction Uncertainty	Single Round Generation	White-box	Yes
Shifting Attention to Relevance (SAR) [30]	Prediction Uncertainty	Single Round Generation	White-box	Yes
P(True) [60]	Prediction Uncertainty	Single Round Generation	White-box	Yes
Response improbability [32]	Prediction Uncertainty	Single Round Generation	White-box	Yes
Average log probability [88]	Prediction Uncertainty	Single Round Generation	White-box	Yes
Predictive Entropy [60]	Prediction Uncertainty	Multi Rounds Generations	White-box	Yes
Relative Mahalanobis distance [112]	Prediction Uncertainty	Multi Rounds Generations	White-box	Yes
HUQ [132]	Prediction Uncertainty	Multi Rounds Generations	White-box	Yes
Conformal Prediction [65, 106]	Prediction Uncertainty	Multi Rounds Generations	White-box	No
ConU [138]	Prediction Uncertainty	Multi Rounds Generations	White-box	No
Level-adaptive conformal prediction [13]	Prediction Uncertainty	Multi Rounds Generations	White-box	No
LoFreeCP [122]	Prediction Uncertainty	Multi Rounds Generations	Black-box	No
Ecc(J),Deg(J) [78]	Prediction Uncertainty	Multi Rounds Generations	Black-box	Yes
Eig(J) [78]	Prediction Uncertainty	Multi Rounds Generations	Black-box	No
Normal length predictive entropy [87]	Prediction Uncertainty	Multi Rounds Generations +Additional Model	White-box	Yes
Semantic Entropy [63]	Prediction Uncertainty	Multi Rounds Generations + Additional Model	White-box	Yes
Kernel Semantic Entropy [101]	Prediction Uncertainty	Multi Rounds Generations + Additional Model	White-box	Yes
Ecc(C),Ecc(E),Deg(C),Deg(E) [78]	Prediction Uncertainty	Multi Rounds Generations + Additional Model	Black-box	Yes
Eig(C),Eig(E) [78]	Prediction Uncertainty	Multi Rounds Generations + Additional Model	Black-box	No
MD-UQ [10]	Prediction Uncertainty	Multi Rounds Generations + Additional Model	Black-box	No
D-UE [17]	Prediction Uncertainty	Multi Rounds Generations + Additional Model	Black-box	Yes

Table 2: An overview of uncertainty quantification methods discussed in this paper.

Maximum token log-probability 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.

Entropy reflects how widely distributed a model's predictions are for a given input, indicating the level of uncertainty in its outputs [60, 63, 86].Entropy for the i-th token is provided by:

$$\mathcal{H}_i = -\sum_{\tilde{w} \in \mathcal{D}} p_i(\tilde{w}) \log p_i(\tilde{w}) \tag{5}$$

Then it is possible to use the mean or maximum value of entropy as the final uncertainty [88]:

$$Avg(\mathcal{H}) = \frac{1}{N}\sum_{i=1}^{N} \mathcal{H}_i; Max(\mathcal{H}) = \max_i(\mathcal{H}_i) \tag{6}$$

Furthermore, Shifting Attention to Relevance (SAR), proposed by Duan et al. [30], 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:

$$R(w_i, s, x) = 1 - |g(x \cup s, x \cup s \setminus \{w_i\})|, \tag{7}$$

Where g is a function that measures the semantic similarity between two sentences such as NLI models.

Response Improbability. Fadeeva et al. [32] 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:

$$MP(s) = 1 - \prod_{i=1} p_i(w_i). \tag{8}$$

If the sentence is very certain (i.e., the product of token probabilities is high), $MP(s)$ will be low.

P(True) Kadavath et al. [60] proposes P(True)[2], which measures the uncertainty of the claim by asking the LLM itself whether the generation is true or not. Specifically, P(True) is calculated:

$$\text{P(True)} = 1 - p(y_1 = \text{“True”}). \tag{9}$$

Note that here we are using y_1 as the first token instead of w_1 because w_1 represents the first token to the generation s while y_1 represented the first token when asking LLM whether the generation s is correct or not. P(True) requires to run the LLM twice. However, it does not require multiple generations s. Therefore, we still classify this method to single-round generation. (This could be considered an uncertainty estimate as the sequence to be evaluated is the prediction given the input.)

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.

[2]The original name is P(IK), which stands for "I Know".

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 [60] can also be applied to multiple response settings and shows a better uncertainty quality based on multiple outputs variability. Similarly, SAR [30] could also be applied to multiple responses. Another study refines this method with Monte Carlo-based approximations [87]. It focuses on how probability distributions evolve across tokens during autoregressive generation. There are two main approaches: one that directly estimates uncertainty by averaging entropy across multiple sampled outputs and another that decomposes sequence-level uncertainty into token-level contributions using a structured entropy approximation.

Conformal Prediction. Conformal Prediction (CP) is a statistical framework that provides formal coverage guarantees for uncertainty estimation in LLMs. It has been widely adopted in recent research due to its distribution-free properties, making 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 [122], using frequency-based sampling combined with normalized entropy (NE) and semantic similarity (SS) to define nonconformity scores. Another black-box CP method introduces a self-consistency-based uncertainty measure [138], 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 [106], 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 [65], using model confidence scores to calibrate prediction sets, ensuring coverage with minimal set size. A more advanced technique, conditional CP [13], 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. One approach leverages Jaccard similarity [78], which measures the overlap between words in different generations. This method evaluates the deviation from self-consistency, where a high Jaccard similarity across generations implies low uncertainty.

However, in natural language generation (NLG) tasks, word-level similarity alone is insufficient, as different responses can convey the same meaning using different phrasing. 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 extend multiple rounds of generation by incorporating external models to assess the consistency of generated responses beyond lexical similarity.

Semantic Entropy. Semantic Entropy (SE) [63], refines uncertainty estimation by clustering generated responses based on semantic equivalence. This approach uses a Natural Language Inference (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. Another method, Kernel Language Entropy (KLE), applies a kernel-based framework to quantify semantic uncertainty [101]. To enhance the performance of semantic entropy, KLE represents them in a semantic space using positive semidefinite kernels. By computing von Neumann entropy over these response distributions, KLE provides an even more fine-grained measure of uncertainty that considers nuanced semantic variations.

Semantic Similarity. Semantic similarity uncertainty methods use Natural Language Inference (NLI) models to measure the semantic relationships between multiple generated responses from an LLM [78]. Instead of relying on lexical overlap, these approaches construct a similarity matrix based on entailment and contradiction scores between generated outputs. A confident model produces responses with high internal consistency, while greater semantic dispersion in the similarity matrix indicates higher uncertainty.

To quantify this dispersion, graph-based spectral metrics are applied. Eccentricity (Ecc) measures the spread of response variability, eigenvalue-based measures (Eig) analyze the spectral properties of the similarity matrix to detect uncertainty, and degree (Deg) evaluates response connectivity, with higher degrees indicating stronger confidence. Chen et al. [10] proposes MD-UQ, which enhanced the performance of graph-based spectral metrics by using an additional knowledge dimension input extracted from an auxiliary LLM and tensor decomposition while Da et al. [17] also enhanced the performance by analyzing the directed graph instead of undirected graph originally from Lin et al. [78].

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. Reading comprehension benchmarks include CoQA for conversational Q&A, RACE for general reading comprehension, TriviaQA for fact-based questions, CosmosQA for contextual understanding, SQuAD for question-answering on passages, and HotpotQA for multi-hop reasoning. Reasoning and math benchmarks include HotpotQA and StrategyQA, which test multi-hop reasoning, GSM8K for solving math problems, and CalibratedMath, designed to evaluate confidence expression in arithmetic. Factuality evaluation draws on datasets such as TruthfulQA for addressing common misconceptions, FEVER for claim verification, and HaluEval for detecting hallucinations, and annotated FActScore dataset for evaluating the factuality of long-form text

Category	Benchmarks
Reading Comprehension	TriviaQA [59], CoQA [110], RACE [66], CosmosQA [47], SQuAD [107], HotpotQA [144]
Reasoning & Math	StrategyQA [36], HotpotQA [144], GSM8K [14], CalibratedMath [76]
Factuality	TruthfulQA [77], FEVER [124], HaluEval [71], FActScore [91]
General Knowledge	MMLU [42], GPQA [111], HellaSwag [151]
Consistency & Ambiguity	ParaRel [31], AmbigQA [93], AmbigInst [44]

Table 3: Categorization of datasets and benchmarks.

generated by LLMs [154]. Several general knowledge benchmarks have been adapted for uncertainty quantification, such as MMLU for a wide range of subjects, GPQA for multiple-choice questions in physical sciences, and HellaSwag for common-sense reasoning through sentence completion. These benchmarks can be adapted for UQ because the task reduces to a binary 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.

Additional categories include consistency benchmarks such as ParaRel, which tests semantic consistency across 328 paraphrases for 38 relations, and datasets like AmbigQA and AmbigInst, which feature inherent ambiguities. Ambiguity datasets are useful in uncertainty quantification (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 introduce dedicated uncertainty quantification (UQ) benchmarks for large language models (LLMs). Yang et al. [143] introduced MAQA, a dataset specifically designed to evaluate epistemic uncertainty in language models. Fadeeva et al. [33] created LM-Polygraph, which was later adopted as a comprehensive uncertainty benchmark by Vashurin et al. [130]. Ye et al. [147] developed a benchmark using conformal prediction across five common NLP tasks. These contributions represent specialized datasets explicitly designed to assess UQ capabilities in LLMs, rather than adapting existing general-purpose benchmarks. Overall, we show the categorization of datasets and benchmarks for UQ in Table 3.

4.2 Evaluation Metrics

Uncertainty quantification (UQ) is often evaluated as a binary classification task, 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 uncertainty score discriminates correct from incorrect predictions [8, 63, 78, 81, 140].

Similarly, **AUPRC** (Area Under the Precision-Recall Curve) and **AUARC** (Area Under the Accuracy-Rejection Curve) [100] also offer further insights into uncertainty quantification. AUPRC measures how well the uncertainty score separates correct from incorrect responses [80], while AUARC assesses how effectively the uncertainty measure aids in selecting accurate responses by determining which uncertain questions to reject [79].

In the context of NLG where the correctness label is hard to obtain, researchers also compute heuristic-based (fuzzy matching) metrics such as **BLEU** [104] and **ROUGE** [63] 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 [63, 154].

Apart from the binary classification framework, there are also multiple evaluation methods designed for specific treatment of uncertainty, sometimes qualitative. For example, focusing on decomposing aleatoric and epistemic uncertainty, Hou et al. [44] evaluates only the aleatoric part by using AmbigQA [93], as high ambiguity questions should incur higher aleatoric uncertainty (whereas math questions, for examples, might have lower). The evaluation in Giulianelli et al. [37], 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, Zhang et al. [154] compare the uncertainty estimate against FActScore [91], as the "correctness" of a long paragraph could be ill-defined or ambiguous.

5 Applications in LLMs

LLMs are increasingly applied in diverse domains, offering flexibility and reasoning capabilities [23]. However, uncertainty quantification (UQ) is crucial for ensuring their reliability, particularly in high-stakes applications like Education [69] and healthcare [24, 115]. This section will introduce the applications that integrate the UQ of LLMs from different domains.

Robotics. LLM-based robotic planning suffers from ambiguity and hallucinations, motivating the need for uncertainty quantification in the planning loop. For example, closed-loop planners [166] employs an uncertainty-based failure detector to continuously assess and adjust plans in real-time, while non-parametric UQ method [126] use an efficient querying strategy to improve reliability. In addition, 'LAP' [97] integrates action feasibility checks to align the LLM's confidence with real-world constraints, improving success rates from approximately 70% to 80%. Similarly, adaptive skill selection [102] dynamically adjusts thresholds for alternative paths, achieving higher success rates in the ALFRED domain. Finally, introspective planning [74] enables LLMs to self-assess uncertainty, enhancing safety and human-robot collaboration.

Transportation. Preliminary research explores how LLMs can enhance transportation systems [19]. For example, LLM inference has been used to bridge the sim-to-real gap in Traffic Signal Control policies [18, 21, 22], and LLM agents have been shown to help smooth mixed-autonomy traffic [145]. However, both cases reveal the potential risk posed by hallucination. A few works have investigated the uncertainty measure while using the LLMs as in [25]; it tries to link the use of visual language models (VLMs) with deep probabilistic programming for uncertainty quantification while conducting multimodal traffic accident forecasting tasks and [17] shows a general solution by modeling the responses as a directed-graph and captures the response uncertainty. Further, UQ exploration anchored on traffic-related LLM use is expected.

Healthcare. In healthcare, LLMs (VLMs) can be good reference for diagnosis, but uncertainty is a critical dimension that should be output together with the generation for more reliable treatment plans. Such as [114] evaluates the ability of uncertainty metrics to quantify LLM confidence when performing diagnosis and treatment selection tasks by assessing the properties of discrimination and calibration. [11] first quantify uncertainty in white-box models with access to model parameters and output logits, and by this, it reveals that an effective reduction of model uncertainty can be achieved by using the proposed multi-tasking and ensemble methods in EHRs. With these findings, however, as [139] benchmarks popular uncertain quantification methods with different model sizes on medical question-answering datasets, authors find that the challenge of UQ for medical applications is still severe.

Education. Many LLM-related products are popular to guide the thinking of students in AI for Education research [160]. However, it is critical to ensure the trustworthiness of LLMs if to participate in a person's thinking logic learning process [109]. The work in [127] discusses the solutions of black-box models UQ evaluation, including Graph Laplacian Eigenvalue Sum, Lexical Similarity, and their reliability. [20] proposes a method that elicits the reasoning topology of the LLM answering process, which can guide the step-by-step analysis of the LLM's reasoning steps before applying it to the student reference guide.

Conclusion. There are many different fields that discuss the need for uncertainty quantification during the LLMs execution process, and the above-discussed ones are only major aspects for now, the multi-agent energy management, operation research, etc, all employ LLMs and would require such consideration respectively.

6 Challenges and Future Directions

While significant strides have been made in integrating uncertainty quantification into Large Language Models, 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 and trustworthiness 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 [70]), yet yield marginal reliability gains (≤ 0.02 AUROC improvement [141]). Hybrid approaches combining low-cost proxies (attention variance [43], hidden state clustering [101]) 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 ("Explain quantum gravity"), knowledge gaps, or decoding stochasticity—a critical barrier in high-stakes domains. Modular architectures that decouple uncertainty estimation layers [46, 117] or employ causal tracing of transformer attention pathways [134] could clarify uncertainty origins. For instance, perturbing model weights [34] might reveal parametric uncertainty in low-resource languages, while input modules flag underspecified terms ("this country") 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 [158], causing 63% of errors in multi-modal QA [155]. Dynamic contrastive decoding and uncertainty-aware fusion protocols show promise[52, 123], but require domain-specific adaptations (e.g., aligning radiology images with reports [73, 161]). Future work must develop unified uncertainty embeddings to harmonize modality confidence scales and adversarial training against cross-modal backdoor attacks [75, 133, 159].

Interventions for Uncertainty. Real-time uncertainty monitoring enables self-correction (e.g., retracting erroneous reasoning steps [148]), but adds 300 to 500ms latency per decoding iteration [141]. Lightweight monitors via attention variance distillation [39, 49] and human-in-the-loop calibration [136] could mitigate overhead while addressing security risks like adversarial uncertainty manipulation (100% attack success rates [152]). Certified defense mechanisms and causal propagation tracing [40] are critical for deploying uncertainty-guided interventions in sensitive applications.

UQ Evaluation. There are several challenges to effectively evaluating UQ quality. There are limitations in adopting a binary classification evaluation approach: It is challenging to decide if a free-form generation is correct even in a QA dataset, let alone more open-ended tasks. Also, even LLM-as-a-judge suffers systematic biases including self-preference bias (where an LLM judge may favor answers written in a style similar to its own) [103], length bias (where LLM judges often prefer longer, more detailed answers) [77], and position bias (where the order of presented answers influences the judgment) [164]. Moreover, current evaluation methods, such as AUROC and AUARC, fall short in capturing nuanced, "meaningful" uncertainty. For instance, they often do not differentiate between cases where a model is confidently wrong (hallucinations) and cases where it expresses a prudent level of uncertainty. Finally, a significant practical challenge in evaluating UQ for LLMs is the lack of suitable datasets that align with what we want to measure. Many benchmarks exist to test LLM capabilities (from reading comprehension to factual QA), but few are explicitly designed for UQ.

7 Conclusion

In this survey, we offer a comprehensive overview of uncertainty quantification (UQ) in Large Language Models (LLMs). Initially, we 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. In terms of methodologies, we systematically evaluate UQ methods using our novel taxonomy, thoroughly 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 a more seamless integration of UQ techniques into LLM development, motivating both machine learning researchers and practitioners to delve into this rapidly advancing area.

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] Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403* (2023).

[3] Amos Azaria and Tom Mitchell. 2023. The Internal State of an LLM Knows When It's Lying. (Dec. 2023), 967–976. doi:10.18653/v1/2023.findings-emnlp.68

[4] 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.

[5] Oleksandr Balabanov and Hampus Linander. 2024. Uncertainty quantification in fine-tuned LLMs using LoRA ensembles. *arXiv preprint arXiv:2402.12264* (2024).

[6] Evan Becker and Stefano Soatto. 2024. Cycles of thought: Measuring llm confidence through stable explanations. *arXiv preprint arXiv:2406.03441* (2024).

[7] Mohammad Beigi, Sijia Wang, Ying Shen, Zihao Lin, Adithya Kulkarni, Jianfeng He, Feng Chen, Ming Jin, Jin-Hee Cho, Dawei Zhou, et al. 2024. Rethinking the Uncertainty: A Critical Review and Analysis in the Era of Large Language Models. *arXiv preprint arXiv:2410.20199* (2024).

[8] 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)*. 5186–5200.

[9] Tiejin Chen, Kuan-Ru Liou, Mithun Shivakoti, Aaryan Gaur, Pragya Kumari, Meiqi Guo, and Hua Wei. [n. d.]. 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*.

[10] 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).

[11] 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).

[12] 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.

[13] 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.

[14] 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).

[15] 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.

[16] Charles Corbière, Nicolas THOME, Avner Bar-Hen, Matthieu Cord, and Patrick Pérez. 2019. Addressing Failure Prediction by Learning Model Confidence. In *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.), Vol. 32. Curran Associates, Inc.

[17] 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).

[18] 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.

[19] 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.

[20] 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).

[21] Longchao Da, Hao Mei, Romir Sharma, and Hua Wei. 2023. Sim2Real Transfer for Traffic Signal Control. In *2023 IEEE 19th International Conference on Automation Science and Engineering (CASE)*. IEEE, 1–2.

[22] 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)*. IEEE, 1124–1129.

[23] Longchao Da, Justin Turnau, Thirulogasankar Pranav Kutralingam, Alvaro Velasquez, Paulo Shakarian, and Hua Wei. 2025. A Survey of Sim-to-Real Methods in RL: Progress, Prospects and Challenges with Foundation Models. *arXiv preprint arXiv:2502.13187* (2025).

[24] 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).

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

[26] Morris H. Degroot and Stephen E. Fienberg. 1983. The Comparison and Evaluation of Forecasters. *The Statistician* 32 (1983), 12–22.

[27] 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.

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

[29] Jakub Drápal, Hannes Westermann, Jaromir Savelka, et al. 2023. Using Large Language Models to Support Thematic Analysis in Empirical Legal Studies.. In *JURIX*. 197–206.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] Xiang Gao, Jiaxin Zhang, Lalla Mouatadid, and Kamalika Das. 2024. Spuq: Perturbation-based uncertainty quantification for large language models. *arXiv preprint arXiv:2403.02509* (2024).

[36] 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* 9 (2021), 346–361. https://api.semanticscholar.org/CorpusID:230799347

[37] 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.

[38] Ethan Goh, Robert Gallo, Jason Hom, Eric Strong, Yingjie Weng, Hannah Kerman, Joséphine A Cool, Zahir Kanjee, Andrew S Parsons, Neera Ahuja, et al. 2024. Large language model influence on diagnostic reasoning: a randomized clinical trial. *JAMA Network Open* 7, 10 (2024), e2440969–e2440969.

[39] Xuan Gong, Abhishek Sharma, Srikrishna Karanam, Ziyan Wu, Terrence Chen, David Doermann, and Arun Innanje. 2021. Ensemble attention distillation for privacy-preserving federated learning. In *Proceedings of the IEEE/CVF international conference on computer vision*. 15076–15086.

[40] Shreya Goyal, Sumanth Doddapaneni, Mitesh M Khapra, and Balaraman Ravindran. 2023. A survey of adversarial defenses and robustness in nlp. *Comput. Surveys* 55, 14s (2023), 1–39.

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

[42] 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).

[43] 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).

[44] 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*. https://openreview.net/forum?id=byxXa99PtF

[45] 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).

[46] 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).

[47] 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).

[48] 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.

[49] Tao Huang, Shan You, Fei Wang, Chen Qian, and Chang Xu. 2022. Knowledge distillation from a stronger teacher. *Advances in Neural Information Processing Systems* 35 (2022), 33716–33727.

[50] 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*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 13441–13460. doi:10.18653/v1/2024.findings-emnlp.785

[51] 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.

[52] 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).

[53] 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*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 2078–2092. doi:10.18653/v1/2020.acl-main.188

[54] Thanmay Jayakumar, Fauzan Farooqui, and Luqman Farooqui. 2023. Large Language Models are legal but they are not: Making the case for a powerful LegalLLM. *arXiv preprint arXiv:2311.08890* (2023).

[55] 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*.

[56] 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* 9 (2021), 962–977. doi:10.1162/tacl_a_00407

[57] 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.

[58] Ming Jin, Qingsong Wen, Yuxuan Liang, Chaoli Zhang, Siqiao Xue, Xue Wang, James Zhang, Yi Wang, Haifeng Chen, Xiaoli Li, et al. 2023. Large models for time series and spatio-temporal data: A survey and outlook. *arXiv preprint arXiv:2310.10196* (2023).

[59] 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 (Volume 1: Long Papers)*. 1601–1611.

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

[61] 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)*, Raúl Vázquez, Hande Celikkanat, Dennis Ulmer, Jörg Tiedemann, Swabha Swayamdipta, Wilker Aziz, Barbara Plank, Joris Baan, and Marie-Catherine de Marneffe (Eds.). Association for Computational Linguistics, St Julians, Malta, 1–14. https://aclanthology.org/2024.uncertainlp-1.1/

[62] Ranim Khojah, Alexander Berman, and Staffan Larsson. 2022. Evaluating N-best Calibration of Natural Language Understanding for Dialogue Systems. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, Oliver Lemon, Dilek Hakkani-Tur, Junyi Jessy Li, Arash Ashrafzadeh, Daniel Hernández Garcia, Malihe Alikhani, David Vandyke, and Ondřej Dušek (Eds.). Association for Computational Linguistics, Edinburgh, UK, 582–594. doi:10.18653/v1/2022.sigdial-1.54

[63] 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*. https://openreview.net/forum?id=VD-AYtP0dve

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

[65] 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).

[66] 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.

[67] Siqi Lai, Zhao Xu, Weijia Zhang, Hao Liu, and Hui Xiong. 2023. LLMLight: Large Language Models as Traffic Signal Control Agents. *arXiv preprint arXiv:2312.16044* (2023).

[68] 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).

[69] Jinsook Lee, Yann Hicke, Renzhe Yu, Christopher Brooks, and René F Kizilcec. 2024. The life cycle of large language models in education: A framework for understanding sources of bias. *British Journal of Educational Technology* 55, 5 (2024), 1982–2002.

[70] 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).

[71] 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.

[72] Xiang Lisa Li, Urvashi Khandelwal, and Kelvin Guu. 2024. Few-shot recalibration of language models. *arXiv preprint arXiv:2403.18286* (2024).

[73] 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.

[74] 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.

[75] 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).

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

[77] 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.

[78] 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).

[79] Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2024. Contextualized Sequence Likelihood: Enhanced Confidence Scores for Natural Language Generation. *arXiv preprint arXiv:2406.01806* (2024).

[80] 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)*, Kevin Duh, Helena Gomez, and Steven Bethard (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 3357–3370. doi:10.18653/v1/2024.naacl-long.184

[81] 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).

[82] 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.

[83] 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*. https://openreview.net/forum?id=jH67LHVOIO

[84] 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).

[85] 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).

[86] Andrey Malinin and Mark Gales. 2018. Predictive uncertainty estimation via prior networks. *Advances in neural information processing systems* 31 (2018).

[87] Andrey Malinin and Mark Gales. 2021. Uncertainty Estimation in Autoregressive Structured Prediction. In *International Conference on Learning Representations*. https://openreview.net/forum?id=jN5y-zb5Q7m

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

[89] 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: EMNLP 2023*. 5011–5034.

[90] Sabrina J. Mielke, Arthur Szlam, Emily Dinan, and Y-Lan Boureau. 2022. Reducing Conversational Agents' Overconfidence Through Linguistic Calibration. *Transactions of the Association for Computational Linguistics* 10 (2022), 857–872. doi:10.1162/tacl_a_00494

[91] Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, 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*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 12076–12100. doi:10.18653/v1/2023.emnlp-main.741

[92] Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. AmbigQA: Answering Ambiguous Open-domain Questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 5783–5797. doi:10.18653/v1/2020.emnlp-main.466

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

[94] 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.

[95] 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.

[96] 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.

[97] 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).

[98] 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.

[99] Allan H. Murphy and Robert L. Winkler. 1984. Probability Forecasting in Meteorology. *J. Amer. Statist. Assoc.* 79 (1984), 489–500.

[100] 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*. PMLR, 65–81.

[101] 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.

[102] 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)*. IEEE, 677–682.

[103] 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.

[104] 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.

[105] 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.

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

[107] 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).

[108] Vipula Rawte, Amit Sheth, and Amitava Das. 2023. A survey of hallucination in large foundation models. *arXiv preprint arXiv:2309.05922* (2023).

[109] Mahefa Abel Razafinirina, William Germain Dimbisoa, and Thomas Mahatody. 2024. Pedagogical alignment of large language models (llm) for personalized learning: a survey, trends and challenges. *Journal of Intelligent Learning Systems and Applications* 16, 4 (2024), 448–480.

[110] 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.

[111] 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*.

[112] 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*. https://openreview.net/forum?id=kJUS5nD0vPB

[113] 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 (Proceedings of Machine Learning Research, Vol. 239)*, Javier Antorán, Arno Blaas, Kelly Buchanan, Fan Feng, Vincent Fortuin, Sahra Ghalebikesabi, Andreas Kriegler, Ian Mason, David Rohde, Francisco J. R. Ruiz, Tobias Uelwer, Yubin Xie, and Rui Yang (Eds.). PMLR, 49–64. https://proceedings.mlr.press/v239/ren23a.html

[114] 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.

[115] Thomas Savage, John Wang, Robert Gallo, Abdessalem Boukil, Vishwesh Patel, Seyed Amir Ahmad Safavi-Naini, Ali Soroush, and Jonathan H Chen. 2025. Large language model uncertainty proxies: discrimination and calibration for medical diagnosis and treatment. *Journal of the American Medical Informatics Association* 32, 1 (2025), 139–149.

[116] 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.

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

[118] 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).

[119] Vaishnavi Shrivastava, Ananya Kumar, and Percy Liang. 2025. Language Models Prefer What They Know: Relative Confidence Estimation via Confidence Preferences. *arXiv preprint arXiv:2502.01126* (2025).

[120] Aviv Slobodkin, Omer Goldman, Avi Caciularu, Ido Dagan, and Shauli Ravfogel. 2023. The curious case of hallucinatory (un) answerability: Finding truths in the hidden states of over-confident large language models. *arXiv preprint arXiv:2310.11877* (2023).

[121] 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. doi:10.1162/tacl_a_00598

[122] 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.

[123] 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).

[124] 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).

[125] 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*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 5433–5442. doi:10.18653/v1/2023.emnlp-main.330

[126] 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).

[127] Dafni Tziakouri. 2024. Education with language models: analyzing uncertainty estimation techniques. (2024).

[128] Ehsan Ullah, Anil Parwani, Mirza Mansoor Baig, and Rajendra Singh. 2024. Challenges and barriers of using large language models (LLM) such as ChatGPT for diagnostic medicine with a focus on digital pathology–a recent scoping review. *Diagnostic pathology* 19, 1 (2024), 43.

[129] 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)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 15440–15459. doi:10.18653/v1/2024.acl-long.824

[130] Roman Vashurin, Ekaterina Fadeeva, Artem Vazhentsev, Lyudmila Rvanova, Akim Tsvigun, Daniil Vasilev, Rui Xing, Abdelrahman Boda Sadallah, Kirill Grishchenkov, Sergey Petrakov, et al. 2024. Benchmarking uncertainty quantification methods for large language models with lm-polygraph. *arXiv preprint arXiv:2406.15627* (2024).

[131] 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).

[132] 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.

[133] Matthew Walmer, Karan Sikka, Indranil Sur, Abhinav Shrivastava, and Susmit Jha. 2022. Dual-key multimodal backdoors for visual question answering. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*. 15375–15385.

[134] 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.

[135] 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*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 3070–3079. doi:10.18653/v1/2020.acl-main.278

[136] Yifan Wang, Jianbin Guo, Shengkui Zeng, Qirui Mao, Zhenping Lu, and Zengkai Wang. 2022. Human-Machine Trust and Calibration Based on Human-in-the-Loop Experiment. In *2022 4th International Conference on System Reliability and Safety Engineering (SRSE)*. IEEE, 476–481.

[137] 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.

[138] Zhiyuan Wang, Jinhao Duan, Lu Cheng, Yue Zhang, Qingni Wang, Xiaoshuang Shi, Kaidi Xu, Hengtao Shen, and Xiaofeng Zhu. 2024. Conu: Conformal uncertainty in large language models with correctness coverage guarantees. *arXiv preprint arXiv:2407.00499* (2024).

[139] 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).

[140] 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*. https://openreview.net/forum?id=gjeQKFxFpZ

[141] 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*.

[142] Adam X Yang, Maxime Robeyns, Xi Wang, and Laurence Aitchison. 2023. Bayesian low-rank adaptation for large language models. *arXiv preprint arXiv:2308.13111* (2023).

[143] Yongjin Yang, Haneul Yoo, and Hwaran Lee. 2024. MAQA: Evaluating Uncertainty Quantification in LLMs Regarding Data Uncertainty. *arXiv preprint arXiv:2408.06816* (2024).

[144] 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).

[145] Huaiyuan Yao, Longchao Da, Vishnu Nandam, Justin Turnau, Zhiwei Liu, Linsey Pang, and Hua Wei. 2024. CoMAL: Collaborative Multi-Agent Large Language Models for Mixed-Autonomy Traffic. *arXiv preprint arXiv:2410.14368* (2024).

[146] 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.

[147] 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* 37 (2025), 15356–15385.

[148] 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.

[149] Spencer Young, Porter Jenkins, Lonchao Da, Jeff Dotson, and Hua Wei. 2024. Flexible heteroscedastic count regression with deep double poisson networks. *arXiv preprint arXiv:2406.09262* (2024).

[150] 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.

[151] 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.

[152] Qingcheng Zeng, Mingyu Jin, Qinkai Yu, Zhenting Wang, Wenyue Hua, Zihao Zhou, Guangyan Sun, Yanda Meng, Shiqing Ma, Qifan Wang, et al. 2024. Uncertainty is fragile: Manipulating uncertainty in large language models. *arXiv preprint arXiv:2407.11282* (2024).

[153] 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).

[154] 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.

[155] 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).

[156] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675* (2019).

[157] Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023. Siren's song in the ai ocean: A survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219* 2, 5 (2023).

[158] Yuan Zhang, Fei Xiao, Tao Huang, Chun-Kai Fan, Hongyuan Dong, Jiawen Li, Jiacong Wang, Kuan Cheng, Shanghang Zhang, and Haoyuan Guo. 2024. Unveiling the tapestry of consistency in large vision-language models. *arXiv preprint arXiv:2405.14156* (2024).

[159] 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).

[160] Zheyuan Zhang, Daniel Zhang-Li, Jifan Yu, Linlu Gong, Jinchang Zhou, Zhanxin Hao, Jianxiao Jiang, Jie Cao, Huiqin Liu, Zhiyuan Liu, et al. 2024. Simulating classroom education with llm-empowered agents. *arXiv preprint arXiv:2406.19226* (2024).

[161] 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.

[162] 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*. https://openreview.net/forum?id=0qSOodKmJaN

[163] Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate Before Use: Improving Few-shot Performance of Language Models. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 12697–12706. https://proceedings.mlr.press/v139/zhao21c.html

[164] 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).

[165] Ou Zheng, Mohamed Abdel-Aty, Dongdong Wang, Zijin Wang, and Shengxuan Ding. 2023. ChatGPT is on the horizon: Could a large language model be all we

need for Intelligent Transportation. *arXiv preprint arXiv:2303.05382* 3, 4 (2023), 5.

[166] 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).