RETHINKING THE UNCERTAINTY: A CRITICAL REVIEW AND ANALYSIS IN THE ERA OF LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

In recent years, Large Language Models (LLMs) have become fundamental to a broad spectrum of artificial intelligence applications. As the use of LLMs expands, precisely estimating the uncertainty in their predictions has become crucial. Current methods often struggle to accurately identify, measure, and address the true uncertainty, with many focusing primarily on estimating model confidence. This discrepancy is largely due to an incomplete understanding of where, when, and how uncertainties are injected into models. This paper introduces a comprehensive framework specifically designed to identify and understand the types and sources of uncertainty, aligned with the unique characteristics of LLMs. Our framework enhances the understanding of the diverse landscape of uncertainties by systematically categorizing and defining each type, establishing a solid foundation for developing targeted methods that can precisely quantify these uncertainties. We also provide a detailed introduction to key related concepts and examine the limitations of current methods in mission-critical and safety-sensitive applications. The paper concludes with a perspective on future directions aimed at enhancing the reliability and practical adoption of these methods in real-world scenarios.

1 INTRODUCTION

Large Language Models (LLMs) have recently demonstrated remarkable capabilities in various complex reasoning and question-answering tasks (Zhao et al., 2023; Wang et al., 2024c; Liang et al., 2022). However, despite their potential, LLMs still face significant challenges in generating erroneous answers (Ji et al., 2023a; Li et al., 2023a; Huang et al., 2023), which can have serious consequences, particularly in domains where high levels of accuracy and reliability are critical. A key issue undermining trust in LLM outputs is the models' lack of transparency and expressiveness in their decision-making processes (Zhou et al., 2023; Lin et al., 2023; Yin et al., 2023; Xiao & Wang, 2018; Hüllermeier & Waegeman, 2021), where comprehensively understanding and estimating the model's uncertainty plays a vital role. For example, in the medical field, a physician diagnosing a critical condition like cancer would not only require a high predictive accuracy from the model but also need to understand the uncertainty associated with the case (Gawlikowski et al., 2022a).

While the need for quantifying uncertainty in LLMs is widely recognized, there still lacks a consensus on the interpretation of uncertainty in this new context (Gawlikowski et al., 2022a; Mena et al., 2021; Guo et al., 2022; Hüllermeier & Waegeman, 2021; Malinin & Gales, 2018), which in turn further complicates its estimation. Terms such as "***uncertainty***", "***confidence***", and "***reliability***" are often used interchangeably, yet they refer to distinct concepts that require careful distinction (Gawlikowski et al., 2021). For instance, an LLM can exhibit a high-confidence response to an inherently uncertain and unanswerable question. However, this response could be contextually inappropriate or factually incorrect, illustrating that high confidence does not necessarily correspond to low uncertainty (Gawlikowski et al., 2022b). Thus, the first challenge that remains in the literature is to *explicate the definition of uncertainty in the context of LLMs and explore the nuanced differences between these intertwined concepts*.

Traditionally, uncertainty in deep neural networks (DNNs) is categorized into two types: *aleatoric*, arising from data randomness such as sensor noise, and *epistemic*, stemming from limitations in model knowledge due to insufficient data or unmodeled complexities (Gawlikowski et al., 2022a; Mena et al., 2021; Guo et al., 2022; Hüllermeier & Waegeman, 2021; Malinin & Gales, 2018). Although these categories are widely used in deep learning, they do not fully address the unique challenges of LLMs, which include processing complex text data, managing extremely large parameters, and dealing with often inaccessible training data. Furthermore, the entire lifecycle of LLMs—from pre-training through inference—introduces unique uncertainties, as does the interaction between users and these models. Understanding these different sources of uncertainty is critical, particularly from the perspective of making LLMs more interpretable and robust. Achieving this understanding, however, is not possible without an *inclusive and fine-grained framework that systematically identifies and analyzes the various sources of uncertainty in LLMs*.

Recently, numerous studies have been proposed, aiming to estimate the uncertainty in LLMs (Manakul et al., 2023; Beigi et al., 2024; Azaria & Mitchell, 2023a; Kadavath et al., 2022; Kuhn et al., 2023), and can be broadly divided into four main categories based on their underlying mechanisms: logit-based (Lin et al., 2022b; Mielke et al., 2022a; Jiang et al., 2021; Kuhn et al., 2023), self-evaluation (Kadavath et al., 2022; Manakul et al., 2023; Lin et al., 2024a), consistency-based (Portillo Wightman et al., 2023; Wang et al., 2023), and internal-based (Beigi et al., 2024). However, given the unique characteristics and nuanced aspects of uncertainty in LLMs, critical questions arise regarding the effectiveness of each type of method in truly capturing uncertainly or other related aspects in the context of LLMs, and which specific sources of uncertainty are being detected across the various stages of an LLM's lifecycle. Answering these questions is vital for developing more reliable and comprehensive approaches to uncertainty estimation in LLMs.

To address the aforementioned challenges and questions, we conduct a critical review and analysis of studies related to uncertainty and other related concepts, aiming to present a comprehensive survey covering the full spectrum of uncertainty in LLMs, particularly focusing on the interplay between uncertainty concepts, sources, estimation methods, and text data characteristics, which, to the best of our knowledge, is still lacking in this field. In summary, our contributions in this survey are manifold and pioneering: (1) We have standardized definitions for uncertainty and explore related concepts, enhancing communication across the field (Section 2). (2) We are the first to propose a comprehensive framework that analyzes all sources of uncertainty throughout the lifecycle of LLMs, providing deep insights into their origins and effective management strategies (Section 3). (3) We evaluate and compare current methods for estimating and evaluating LLM uncertainty, discussing their strengths and limitations (Section 4). (4) Finally, we identify future research directions to enhance uncertainty estimation in LLMs, addressing critical gaps and emerging trends for improved reliability and accuracy in critical applications (Section 5).

2 Definition of Uncertainty and Related Concepts

This section begins by offering a comprehensive definition of *uncertainty* and its associated concepts—*confidence* and *reliability*—within the context of large language models. As illustrated in Figure 1, although these concepts are interrelated, they pertain to distinct aspects of model performance that necessitate careful differentiation. We specifically emphasize the differences between *uncertainty* and *confidence*, terms that are often used interchangeably in the literature.

Uncertainty fundamentally refers to the extent to which a model "**knows**" or "**does not know**" about a given input, based on the training it has received (Malinin & Gales, 2018; Der Kiureghian & Ditlevsen, 2009; Hüllermeier & Waegeman, 2021; Gawlikowski et al., 2022a; Kendall & Gal, 2017). Often, this arises

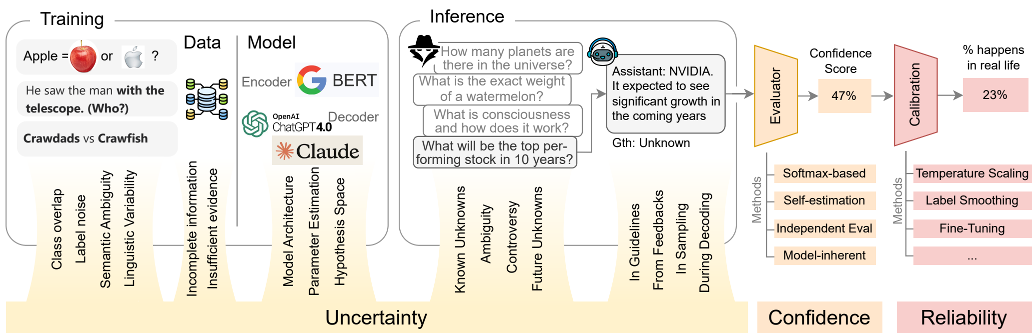

Figure 1: Visualization of the Distinct Aspects of Uncertainty, Confidence, and Reliability in Large Language Models

from inadequate or conflicting training data (Guo et al., 2022; Hüllermeier & Waegeman, 2021; Mena et al., 2021), inappropriate model selection (Gawlikowski et al., 2022a; Mena et al., 2021; Battaglia et al., 2018), or factors like noise and inherent data ambiguity (Kendall & Gal, 2017; Mena et al., 2021; Guo et al., 2022). These factors collectively delineate a model's understanding—or misunderstanding—of its operational environment, influencing the reliability of its outputs.

Confidence, often expressed as a "predicted probability score", quantifies the likelihood that a model's prediction is correct. Derived from the softmax output applied to logits, this score assigns each class a probability between 0 and 1, with the highest probability indicating the model's chosen prediction (He et al., 2024; Guo et al., 2017; Nandy et al., 2021). However, confidence scores can be misleading; they may exhibit "overconfidence", where the score is high despite inaccurate predictions, or "underconfidence", where scores are low even when predictions are correct (Lakshminarayanan et al., 2017; Wang, 2023; Chen et al., 2022; Guo et al., 2017).

Reliability. Merely estimating confidence scores is insufficient for safe decision-making. It's vital to align confidence scores with the actual probabilities of correct predictions, a process known as calibration (Guo et al., 2017; Wang, 2023). Re-calibration techniques have been developed to enhance this alignment, ensuring that confidence estimates are accurately calibrated and reliable for practical applications (Guo et al., 2017; Wang, 2023; Nixon et al., 2019; Mukhoti et al., 2020).

Does High Confidence Score Always Mean Low Uncertainty? Confidence scores are often interpreted as indicators of uncertainty, leading to significant challenges. DNNs frequently assign high confidence to inputs far removed from their training data, resulting in misleading confidence levels for incorrect classifications (Hein et al., 2019). For instance, a network trained on images of cats and dogs might confidently, but incorrectly, classify a bird as one of these categories (Gawlikowski et al., 2022a; Malinin & Gales, 2019). Therefore, high confidence does not necessarily indicate low uncertainty, making such an assumption problematic. LLMs also exhibit high confidence even when uncertainty is substantial. For example, models might confidently answer ambiguous or unanswerable questions like "*How many planets are in the universe?*" (**known unknowns**), "*What will be the top performing stock in 10 years?*" (**future unknowns**), "*What is consciousness and how does it work?*" (**controversial unknowns**), or "*What is the exact weight of a watermelon?*" (**ambiguous questions**), despite the inherent uncertainty of such questions. Similarly, models may express high certainty in speculative or hypothetical scenarios, like "*What would happen if the US had lost the Independence War?*", even when no definitive answer exists. These examples highlight that a high confidence score which derived through various computational methods, often imply as zero uncertainty, does not necessarily indicate the correctness of an answer. It is crucial, therefore, to approach high confidence scores with caution and to develop methods to measure the true uncertainty.

3 Sources of Uncertainty in Large Language Models

3.1 A Comprehensive Framework for Understanding Uncertainty in LLMs

There are three traditional categories of uncertainty commonly used in deep learning, including (1) **Model (epistemic) Uncertainty**, which pertains to uncertainties in estimating model parameters, reflecting model fit and its limitations in generalizing to unseen data (Der Kiureghian & Ditlevsen, 2009; Lahlou et al., 2023; Hüllermeier & Waegeman, 2021; Malinin & Gales, 2018); (2) **Data (or aleoteric) Uncertainty** that stems from complexities within the data itself, such as class overlap and various types of noise (Der Kiureghian & Ditlevsen, 2009; Rahaman & Thiery, 2020; Wang et al., 2019; Malinin & Gales, 2018); and (3) **Distributional Uncertainty**, which often dues to dataset shift and occurs when training and testing data distributions differ, leading to potential generalization issues during real-world applications where the model faces data markedly different from what it was trained on (Malinin & Gales, 2018; Nandy et al., 2021; Gawlikowski et al., 2022a; Chen et al., 2019; Mena et al., 2020).

These traditional uncertainty categories, though prevalent in deep learning, fail to fully address the unique challenges of LLMs. LLMs are characterized by extensive parameters, complex text data processing, and often limited access to training data, introducing specific uncertainties in their outputs. Moreover, interactions with users in dynamic environments and human biases in data annotation or model alignment complicate the uncertainty landscape. Unlike general deep learning models that primarily predict numerical outputs or classes, LLMs generate knowledge-based outputs which may include inconsistent or outdated information (Lin et al., 2024b). These features cannot be adequately addressed by simply categorizing uncertainty into three traditional types. These distinctive aspects necessitate a comprehensive framework to better understand the diverse sources of uncertainty in LLMs.

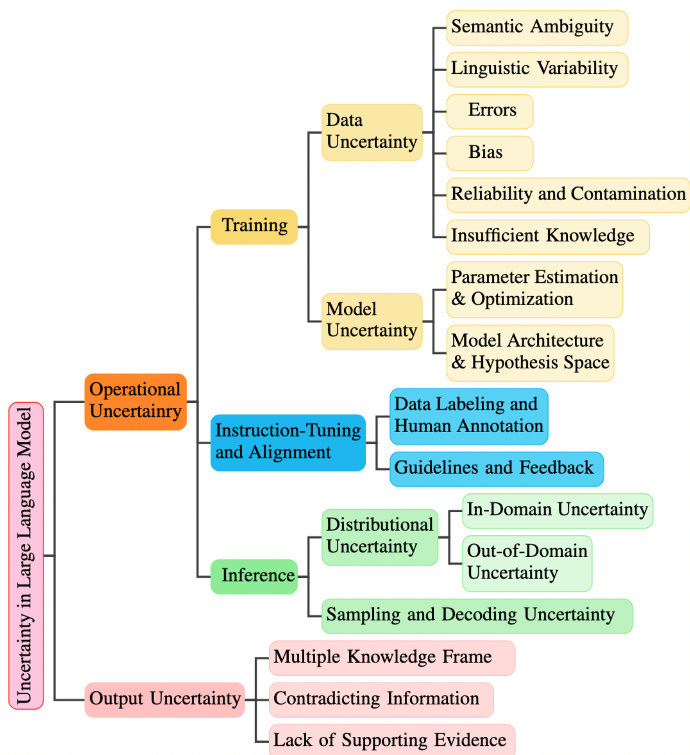

Figure 2: A comprehensive framework to categorize the sources of uncertainty through LLM's life-cycle.

To address these challenges, we introduce a new framework to categorize uncertainty in LLMs, as illustrated in Figure 2. This framework distinguishes between ***operational* uncertainty**, which pertains to model and data processing, and ***output uncertainty***, focusing on the quality of the generated content. Specifically:

Operational uncertainty in LLMs arise from pre-training to inference, encompassing data acquisition, model and architecture design, training and optimization processes, alignment, and inference activities. These uncertainties stem from how LLMs are trained on extensive datasets, process inputs, and generate text. In essence, operational uncertainties arise when the model is unable to capture the full complexity of the data it has been trained on, or when the input data itself introduces ambiguities or noise.

Output uncertainty in LLMs stems from challenges in analyzing and interpreting the generated text, relating to the quality and reliability of information used for decision-making. For example, in medical scenarios, where an LLM is tasked with providing diagnostic suggestions based on patient symptoms, the model may generate multiple possible diagnoses. However, if these suggestions lack justification or supporting evidence, or contain contradictory information, significant uncertainty arises for the physician who must determine the credibility of these diagnoses. The physician may face substantial challenges in deciding which diagnosis to investigate further, highlighting the critical need for LLMs to provide well-supported, consistent, and reliable outputs to ensure their practical utility in decision-making processes.

By distinguishing between operational and output uncertainties, our framework offers several benefits: first, a fine-grained approach that captures the unique features of LLMs, providing a precise reflection of uncertainty for better modeling and understanding. Second, it establishes a foundation for identifying uncertainty sources, essential for developing targeted methods to accurately quantify them. Third, it offers stakeholder-specific insights, helping developers, users, and administrators address uncertainties relevant to their roles, enhancing model robustness, user interaction, and governance. Lastly, by aggregating beliefs and evaluating output evidence, the framework builds trust in LLM outputs, particularly in critical fields like medical diagnosis or legal reasoning.

3.2 Operational Uncertainty in LLMs

3.2.1 Operational Uncertainty in Pre-Training and Training Stage of LLMs

Data Uncertainty This phase involves collecting and organizing the pre-training corpus, which is critical as the quality, diversity, and representativeness of the data directly affect the model's understanding and ability to generalize. The sources of uncertainty at this stage include:

(1) Semantic Ambiguity: Textual data inherently contain semantic complexities, leading to significant uncertainties in training and inference processes for language models. For example, the word 'bank' can mean a financial institution or a river's edge, and 'lead' can refer to the act of leading or the metal, depending on the context (Anand & Kumar, 2022; Ott et al., 2018; Dreyer & Marcu, 2012; Blodgett et al., 2020). These semantic ambiguities pose challenges in maintaining meaning across different contexts and highlight the difficulty of achieving consistent semantic understanding. Such ambiguities are a primary source of uncertainty, complicating the model's understanding (Anand & Kumar, 2022; Piantadosi et al., 2011).

(2) Linguistic Variability: Textual environments are dynamic and subject to contextual and cultural shifts that significantly alter data interpretation and relevance (Kutuzov et al., 2018; Levy, 2008). For example, word meanings and usages evolve, new slang emerges, and topics range from casual conversations to specialized discussions, each with distinct linguistic nuances (Levy, 2008; Liu et al., 2018; Kutuzov et al., 2018). This variability requires language models to continually interpret context to determine meaning, greatly increasing the uncertainty in their knowledge and response.

(3) Errors: The data collection process can introduce errors like typographical mistakes, incorrect tagging, or grammatical errors. These inconsistencies can significantly mislead the learning process, impairing the LLM's ability to model and generate text accurately (Wang et al., 2024a; Liu et al., 2018).

(4) Insufficient Coverage: This refers to situations where incomplete coverage in the training dataset leads to uncertainty. Mitigating this requires acquiring more extensive and diverse data that encompasses various viewpoints (Gawlikowski et al., 2022a).

(5) Reliability and Contamination of Data: Training data for LLMs often contains inaccuracies or outdated content (Lin et al., 2024b), which can lead these models to perpetuate and amplify such errors (Lin et al., 2024b). Misrepresented facts and contaminated data—incorrect or misleading information included in the training set—introduce significant uncertainty and hinder the models' reliability as decision-making tools (Jiang et al., 2024).

(6) Human Biases: In training data related to gender, race, socio-economic status, age, or disability are primary sources of uncertainty in LLM predictions (Bender & Friedman, 2018). These biases skew the model's understanding and responses, resulting in outputs that may not be universally valid or appropriate, thus increasing uncertainty about the model's performance and reliability in diverse real-world scenarios (Kirk et al., 2021).

Model Uncertainty This type primarily arises from the model's fit to the data, highlighting its ability to generalize from the training data to unseen data (Malinin & Gales, 2018). The design of the architecture and the training process of an LLM are crucial in shaping its capabilities and effectiveness. This process involves the strategic configuration of the neural network, where each decision reflects an inductive bias—the underlying assumptions embedded in the model through choices in network structure. These biases influence how the model interprets and processes information (Battaglia et al., 2018). The sources of uncertainty pertain to model uncertainty are:

(1) Model Architecture and Hypothesis Space: The architecture of an LLM and the hypothesis space it explores are crucial to its performance and error susceptibility. Architectural decisions, such as the number of layers and network types, significantly influence model effectiveness across various tasks. These choices determine the hypothesis space—what the model can learn and predict—thereby introducing uncertainty in the model's ability to understand and generate language under different conditions. Variability in architectural setup can cause performance discrepancies when applied to new or varied datasets (Fedus et al., 2022; Abdar et al., 2021b; He et al., 2024; Gawlikowski et al., 2022a; Dodge et al., 2020).

(2) Parameter Estimation and Optimization: Parameter estimation and optimization methods are critical sources of variability and uncertainty in LLMs. Choices in optimization techniques (e.g., SGD, Adam), learning rates, loss functions, and regularization methods (e.g., dropout, L2 regularization) significantly impact the model's generalization capabilities and robustness (Lakshminarayanan et al., 2017). These factors contribute to uncertainty in the model's ability to consistently replicate results across different runs or datasets, affecting its adaptability to new data and stability across various operational environments (Payzan-LeNestour & Bossaerts, 2011).

3.2.2 Operational Uncertainty in Instruction Tuning and Alignment Stage of LLMs

Instruction-tuning and Reinforcement Learning from Human Feedback (RLHF) are advanced techniques that enhance LLMs' adaptability and responsiveness to specific tasks or user preferences (Ouyang et al., 2022; Rafailov et al., 2024). These techniques refine model responses to align more closely with expected outcomes using predefined instructions or guidelines (Bai et al., 2022; Askell et al., 2021). This process introduces uncertainty through two main sources:

(1) Inconsistency and Bias in Data Labeling and Human Annotation: Human labeling and annotation are fundamental sources of uncertainty in the training LLMs (Ghandeharioun et al., 2019; Abdar et al., 2021a; Zhou et al., 2024). The subjective nature of human judgment introduces variability and biases, affecting learning outcomes from reward models (Zhang et al., 2023a). Individual differences in perception and decision-making can lead to cognitive biases and inconsistency in labeled data, which RLHF reward modeling algorithms may exacerbate (Wang et al., 2024b; Denison et al., 2024).

(2) Interpretation of Guidelines and Feedback:The interpretation of instructions can vary, depending on the clarity of the guidelines and the model's ability to interpret them contextually Wang et al. (2024a). Discrepancies in understanding or applying these instructions can lead to variability in the model's outputs (Siththaranjan et al., 2023; Wu et al., 2024; Chidambaram et al., 2024; Park et al., 2024). The subjective nature of feedback and its interpretation by the model can introduce additional layers of uncertainty. This is particularly evident when there is a lack of consensus among human reviewers, leading to challenges in achieving stable and predictable model behavior (Ghandeharioun et al., 2019; Abdar et al., 2021a; Zhou et al., 2024).

3.2.3 Operational Uncertainty in Inference Stage of LLMs

Distributional Uncertainty occurs when there are discrepancies between the training data distributions and those encountered during testing, a phenomenon known as dataset shift. This uncertainty is prevalent in real-world applications where models face data significantly different from their training sets. Distributional uncertainty indicates a lack of model familiarity with new data, posing challenges in making accurate predictions. This uncertainty is categorized into in-domain and out-of-domain types.

(1) In-Domain Uncertainty: This type of uncertainty occurs when LLMs operate within their training environments and inputs closely resemble the training data distribution. Such scenarios often present interpolation challenges or deal with 'unknown knowns' (Ashukha et al., 2021; Hüllermeier & Waegeman, 2021). Despite the similarity to training datasets, subtle nuances and variations within seemingly familiar data can still provoke uncertainties if not fully captured during training (Kim et al., 2023; Kong et al., 2020).

(2) Out-of-Domain Uncertainty: This occurs when LLMs face queries or data points outside their training distribution, leading to 'unknown knowns' and 'unknown unknowns,' where the model lacks the necessary data or precedent to generate well-founded responses, often resulting in overly generic or shallow outputs (Xu & Ding, 2024; Liu et al., 2024; Kong et al., 2020). 'Unknown knowns' are situations where the model has indirect knowledge but encounters data that, while potentially interpolatable from known data, still lies outside its direct experience, leading to uncertain responses despite possible correctness (Amayuelas et al., 2023). Conversely, 'unknown unknowns' refer to entirely unfamiliar data types or topics that the model has never encountered, typically producing speculative, erroneous, or hallucination.

Sampling and Decoding Strategy is another importatn sources of uncertainty in inference stage originate from the . The configuration of LLMs at inference time, including temperature scaling, context length (Anil et al., 2022), and decoding strategies such as beam search or nucleus sampling, significantly affects the uncertainty of model outputs. Temperature scaling adjusts the randomness of predictions by modifying the probability distribution, with lower values resulting in more deterministic outputs and higher values increasing diversity and variability. The choice of context length can influence the extent of uncertainty in outputs, with longer generations potentially introducing more uncertainty than shorter ones. Decoding strategies like beam search enhance output coherence by considering multiple possibilities, yet may reduce variability and creativity. These configurations are crucial for the model's generalization across tasks, impacting the coherence and consistency of predictions, thus playing a key role in balancing performance with uncertainty (Renze & Guven, 2024; Xie et al., 2023; Zeng et al., 2021; Ott et al., 2018; Hashimoto et al., 2024; Stahlberg & Byrne, 2019; Eikema & Aziz, 2020; Meister et al., 2020; Fan et al., 2018; Holtzman et al., 2020; Hewitt et al., 2022).

3.3 Output Uncertainty in LLMs

In contrast to operational uncertainties, which stem from the mechanics of how LLMs as deep neural networks generate text, output uncertainties focus on the outputs these models produce when used as knowledge generation tools. This capability can release a vast flow of information useful for decision-making in various downstream tasks. Here, the challenge shifts from a shortage of information to the risks of poorly understanding and managing inherent uncertainties, which may arise from unreliable, incomplete, deceptive, or conflicting information. The topic of reasoning and decision-making under uncertainty has been extensively explored in various AI domains, such as belief/evidence theory and game theory. This extensive knowledge base is crucial for enhancing our understanding of LLM outputs and identifying their inherent uncertainties, important for tasks relying on this knowledge. Building on insights from a recent survey on uncertainty and belief theory (Guo et al., 2022), we adapted their framework to classify uncertainties better suited to the characteristics of LLM outputs. As a result, we categorize the sources and causes of output uncertainties based on the ***ambiguity*** in the output itself, which can stem from various causes and sources as:

(1) Lack of supporting evidences and incomplete knowledge: An LLM may produce outputs with uncertainties due to a lack of supporting evidence in the responses, a common occurrence even in well-trained models addressing complex or nuanced topics. This uncertainty stems from the model generating conclusions without sufficient evidence to substantiate its answers, and from its inability to provide sufficient theoretical understanding or reliable information. The connections between claims and supporting information may not be clearly established or detailed, which hampers confident and reliable reasoning and decision-making. To mitigate this, the model's output can be enriched by incorporating more robust evidence or discarding unreliable evidence. Additionally, the complexity of model-generated information can overwhelm users due to limited cognitive capacity to process dense or intricate data. Simplifying the data into more manageable chunks with coarser granularity or focusing on key features while omitting less critical details can help. Effectively managing this uncertainty involves concentrating on leveraging the most relevant information available, thus enhancing the confidence and trust in the accuracy and relevance of the outputs.

(2) Multiple knowledge frames and contradicting knowledge : These sources of uncertainty arise in scenarios where the same information—such as evidence or opinions—can be interpreted in various ways, leading to conflicting views. Multiple, valid beliefs about certain knowledge or information may coexist, often due to conflicting evidence. Conflicts can occur when parts of the information are incorrect, irrelevant, or when the model interpreting the data is not suitable for the current context. Additionally, conflicts may arise where there is no definitive ground truth, or in cases of controversial debate. Differing opinions from users, based on subjective perspectives, further complicate understanding and increase the layers of uncertainty.

4 Approaches for Estimating and Evaluating Uncertainty in LLMs

Existing approaches to assess how well a model understands and is certain about its predictions in LLMs can be summarized into four major categories:

Logit-based approaches (Lin et al., 2022b; Mielke et al., 2022a; Jiang et al., 2021; Kuhn et al., 2023) assess model confidence by analyzing the probability distributions or entropy of outputs, providing clear measures of confidence. Although straightforward to implement, a fundamental issue with using logits as confidence indicators is that they reflect the probability distribution across potential tokens (vocabulary space), capturing linguistic forms rather than verifying the truthfulness or correctness of statements (Lin et al., 2022b; Si et al., 2022; Tian et al., 2023). Logit probabilities, irrespective of their magnitude, predominantly represent distribution over vocabulary space. Therefore, logit probabilities do not directly indicate model uncertainty but also reveal various linguistic factors that influence the model's output. This nuanced perspective is in stark contrast to human expressions of certainty, which typically reflect a belief in the accuracy or truth of a claim, based on information processing and decision-making processes, and are not influenced by phrasing (Koriat et al., 1980; Fischhoff et al., 1977). Additionally, another significant limitation is that logit-based methods do not identify or measure any types of uncertainty, limiting their applicability in scenarios where understanding the sources and degrees of uncertainty is crucial.

Consistency-based approaches (Vazhentsev et al., 2023; Portillo Wightman et al., 2023; Wang et al., 2023; Shi et al., 2022; Manakul et al., 2023; Agrawal et al., 2023) assess confidence by evaluating the agreement among various model responses, identifying potential inconsistencies. However, these methods encounter significant challenges, especially due to the diversity of potential paraphrases and formatting variations in textual data, complicating their use in real-time scenarios (Xiong et al., 2024; Jiang et al., 2021; Fadaee et al., 2017; Li et al., 2022; Ding et al., 2024; Kuhn et al., 2023). Additionally, a non-trivial challenge is the effective measurement of consistency among responses, a persisting issue that hinders accurate confidence assessment (Manakul et al., 2023; Zhang et al., 2023b).

Self-evaluation methods (Kadavath et al., 2022; Manakul et al., 2023; Lin et al., 2024a) enable models to internally assess the correctness of their answers by leveraging their introspective capabilities. These methods employ various prompts that encourage models to express their confidence through numerical values or verbalized terms. Recent studies have refined these approaches, utilizing strategies like Chain of Thought (CoT) to enhance how models calibrate and articulate linguistic confidence (Xiong et al., 2023). Research has also explored expressing confidence with linguistics qualifiers to better align verbal expressions with the model's actual confidence levels (Mielke et al., 2022b; Zhou et al., 2023; Lin et al., 2022a), making model outputs more understandable for users. Despite their potential, these methods are constrained by the model's limited self-awareness, which can lead to circular reasoning and overconfident inaccuracies (Ji et al., 2023b; Chen et al., 2023). Another challenge is the interpretability and validity of obtained probabilities that align with specific linguistic and psychological interpretations, including expressions like 'I think,' 'undoubtedly,' or 'high confidence.'

Internal-Based Approach Recently, Beigi et al. (2024) used a mutual information framework to theoretically demonstrate that the internal states of large language models provide additional insights into the correctness of their answers. Burns et al. (2023) introduced an innovative unsupervised method that maps hidden states to probabilities. This approach involves responding to "Yes" or "No" questions, extracting model activations, converting these activations into probabilities. Furthering this research, studies have employed linear probes (Li et al., 2023b; Azaria & Mitchell, 2023b) and contrastive learning (Beigi et al., 2024) to assess whether the internal states across various layers can distinguish between correct and incorrect answers. Empirical results suggest that certain middle layers and specific attention heads show strong discriminative abilities. Beigi et al. (2024) expanded these findings by illustrating that for tasks requiring contextual processing, such as reading comprehension, the outputs of multi-head self-attention (MHSA) components are crucial for assessing response correctness. However, current methodologies exhibit limitations. Each task and dataset requires training a specific "confidence estimator" model, which restricts their generalizability. This limitation is evident as the performance of these methods often declines when trained on one task and dataset and tested on another, highlighting their limited transferability across different applications (Bashkansky et al., 2023). Additionally, the computational resources required to train these confidence estimators pose challenges for their deployment in real-time applications, further complicating their practical utility.

What are the main properties of these methods in estimating uncertainty in LLMs? Table 1 outlines the key characteristics of the methods discussed in this study, including their complexity, computational effort, memory consumption, flexibility, and their ability to identify and measure sources of uncertainty in LLMs.

Description	Logit-Based	Internal-Based	Self-Evaluation	Consistency-Based
Uncertainty/Confidence	Confidence Score	Confidence Score	Confidence Score	Confidence Score
Identifying Sources	No	No	No	No
Need Access to Parameters	Yes	Yes	No	No
Explainability	No	somehow	No	No
Transferability	High	Low	High	Mid
Evaluation Metrics	Acc, ECE	Acc, ECE	Acc, ECE	Acc, ECE
Accuracy	Low	High	Very Low	Low
Need Training?	No	Yes	some methods	No
Comp. Effort Training	Low	High	Mid	Mid
Mem. Consumption Training	Low	High	Low	Mid
Comp. Effort Inference	Low	High	Low	Mid
Mem. Consumption Inference	Low	High	Low	Low

Table 1: An overview of the four general methods presented in this paper. The labels *high* and *low* are given relative to the other approaches and based on the general idea behind them.

5 Discussion and Future Direction

Go beyond Confidence Estimation: As discussed, the literature on uncertainty estimation in LLMs primarily interprets confidence scores as measures of uncertainty. Such a prevalent method oversimplifies the nuanced and complex nature of uncertainty inherent in model predictions, which is crucial for accurate interpretation and reliability of model outputs. The limitations inherent in current methodologies necessitate the development of a more advanced framework for categorizing uncertainty estimation in large language models that surpasses the reliance on simple confidence scores.

Lack of Explainability: Current methods of confidence estimation provide certainty predictions without elucidating the underlying causes of potential uncertainties. While these scores may seem reasonable to human observers, the absence of insight into the sources of uncertainty complicates trust in the model's outputs, particularly in safety-critical contexts where explainability is essential. Current confidence quantification techniques struggle to pinpoint specific weaknesses or improvement areas in the model. Additionally, these methods lack the necessary transparency to clarify the reasons for model uncertainty, whether due to input ambiguity from users, insufficient knowledge, or conflicting information in the training data.

Lack of Ground Truth for Uncertainty Estimation: Current methods for estimating uncertainty are empirically evaluated and assess how accurately they predict the correctness of an answer. However, there is generally no established ground truth for validation, particularly regarding the sources of uncertainty, meaning currently there is no metric and method to determine the contribution of different uncertainties for specific models and tasks.

Lack of Transferability of Uncertainty Estimation Methods Across Different Applications and Datasets: Current uncertainty estimation methods often struggle with adaptability and generalizability when applied to new applications or datasets. Although effective within their specific domains, these methods frequently fail to yield reliable results in different settings due to factors like data distribution differences and unique domain-specific requirements. To overcome these limitations, it is crucial to develop more robust and flexible uncertainty estimation techniques that can adjust to the varied conditions and demands of diverse applications.

Lack of Standardized Evaluation Protocol & Comprehensive Benchmarks for Confidence Estimation: Current methods for evaluating uncertainty estimation methods are typically used to compare uncertainty quantification techniques through metrics like accuracy, calibration, or performance in out-of-distribution detection, using standardized datasets common within the LLM community. Despite this, variations in experimental settings across studies highlight the need for a comprehensive benchmark across tasks and domains to assess their robustness. The absence of a standardized testing protocol poses challenges for researchers from different downstream tasks, making it difficult to identify the most advanced methods or choose a specific sub-field of uncertainty quantification to pursue. This lack of uniformity hinders the direct comparison of emerging techniques and impedes the broader acceptance and integration of established uncertainty quantification methods.

6 Conclusion

In this paper, we have reviewed and analyzed the uncertainty inherent in LLMs. We clarified the definition of uncertainty and related concepts, enhancing understanding across various domains. A comprehensive framework was proposed to categorize and identify sources of uncertainty throughout the lifecycle of LLMs. We also reviewed current approaches in the literature, discussed their challenges and limitations, and highlighted future directions to enhance the practicality of LLMs in real-life applications.

REFERENCES

Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U. Rajendra Acharya, Vladimir Makarenkov, and Saeid Nahavandi. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76:243–297, dec 2021a. doi: 10.1016/j.inffus.2021.05.008. URL `https://doi.org/10.1016%2Fj.inffus.2021.05.008`.

Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U Rajendra Acharya, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76:243–297, 2021b.

Ayush Agrawal, Lester Mackey, and Adam Tauman Kalai. Do language models know when they're hallucinating references? *ArXiv preprint*, abs/2305.18248, 2023. URL `https://arxiv.org/abs/2305.18248`.

Alfonso Amayuelas, Liangming Pan, Wenhu Chen, and William Wang. Knowledge of knowledge: Exploring known-unknowns uncertainty with large language models. *ArXiv preprint*, abs/2305.13712, 2023. URL `https://arxiv.org/abs/2305.13712`.

Sanjay Kumar Anand and Suresh Kumar. Uncertainty analysis in ontology-based knowledge representation. *New Gen. Comput.*, 40(1):339–376, apr 2022. ISSN 0288-3635. doi: 10.1007/s00354-022-00162-6. URL `https://doi.org/10.1007/s00354-022-00162-6`.

Cem Anil, Yuhuai Wu, Anders Andreassen, Aitor Lewkowycz, Vedant Misra, Vinay Ramasesh, Ambrose Slone, Guy Gur-Ari, Ethan Dyer, and Behnam Neyshabur. Exploring length generalization in large language models, 2022. URL `https://arxiv.org/abs/2207.04901`.

Arsenii Ashukha, Alexander Lyzhov, Dmitry Molchanov, and Dmitry Vetrov. Pitfalls of in-domain uncertainty estimation and ensembling in deep learning, 2021. URL `https://arxiv.org/abs/2002.06470`.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, et al. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*, 2021.

Amos Azaria and Tom Mitchell. The internal state of an LLM knows when it's lying. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 967–976, Singapore, December 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.68. URL `https://aclanthology.org/2023.findings-emnlp.68`.

Amos Azaria and Tom Mitchell. The internal state of an llm knows when its lying. *ArXiv preprint*, abs/2304.13734, 2023b. URL `https://arxiv.org/abs/2304.13734`.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *ArXiv preprint*, abs/2204.05862, 2022. URL `https://arxiv.org/abs/2204.05862`.

Naomi Bashkansky, Chloe R Loughridge, and Chuyue Tang. Surely you're lying, mr. model: Improving and analyzing ccs. 2023.

Peter W. Battaglia, Jessica B. Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, Caglar Gulcehre, Francis Song, Andrew Ballard, Justin Gilmer, George Dahl, Ashish Vaswani, Kelsey Allen, Charles Nash, Victoria Langston, Chris Dyer, Nicolas Heess, Daan Wierstra, Pushmeet Kohli, Matt Botvinick, Oriol Vinyals, Yujia Li, and Razvan Pascanu. Relational inductive biases, deep learning, and graph networks, 2018. URL https://arxiv.org/abs/1806.01261.

Mohammad Beigi, Ying Shen, Runing Yang, Zihao Lin, Qifan Wang, Ankith Mohan, Jianfeng He, Ming Jin, Chang-Tien Lu, and Lifu Huang. Internalinspector i^2: Robust confidence estimation in llms through internal states, 2024. URL https://arxiv.org/abs/2406.12053.

Emily M. Bender and Batya Friedman. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604, 2018. doi: 10.1162/tacl_a_00041. URL https://aclanthology.org/Q18-1041.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (technology) is power: A critical survey of "bias" in NLP. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5454–5476, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.485. URL https://aclanthology.org/2020.acl-main.485.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=ETKGuby0hcs.

Jiefeng Chen, Jinsung Yoon, Sayna Ebrahimi, Sercan Arik, Tomas Pfister, and Somesh Jha. Adaptation with self-evaluation to improve selective prediction in LLMs. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 5190–5213, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.345. URL https://aclanthology.org/2023.findings-emnlp.345.

Wenhu Chen, Yilin Shen, Hongxia Jin, and William Wang. A variational dirichlet framework for out-of-distribution detection, 2019. URL https://arxiv.org/abs/1811.07308.

Yangyi Chen, Lifan Yuan, Ganqu Cui, Zhiyuan Liu, and Heng Ji. A close look into the calibration of pre-trained language models. *ArXiv preprint*, abs/2211.00151, 2022. URL https://arxiv.org/abs/2211.00151.

Keertana Chidambaram, Karthik Vinay Seetharaman, and Vasilis Syrgkanis. Direct preference optimization with unobserved preference heterogeneity. *arXiv preprint arXiv:2405.15065*, 2024.

Carson Denison, Monte MacDiarmid, Fazl Barez, David Duvenaud, Shauna Kravec, Samuel Marks, Nicholas Schiefer, Ryan Soklaski, Alex Tamkin, Jared Kaplan, Buck Shlegeris, Samuel R. Bowman, Ethan Perez, and Evan Hubinger. Sycophancy to subterfuge: Investigating reward-tampering in large language models, 2024. URL https://arxiv.org/abs/2406.10162.

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

Bosheng Ding, Chengwei Qin, Ruochen Zhao, Tianze Luo, Xinze Li, Guizhen Chen, Wenhan Xia, Junjie Hu, Anh Tuan Luu, and Shafiq Joty. Data augmentation using llms: Data perspectives, learning paradigms and challenges, 2024.

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping, 2020. URL https://arxiv.org/abs/2002.06305.

Markus Dreyer and Daniel Marcu. Hyter: Meaning-equivalent semantics for translation evaluation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 162–171, 2012.

Bryan Eikema and Wilker Aziz. Is MAP decoding all you need? the inadequacy of the mode in neural machine translation. In Donia Scott, Nuria Bel, and Chengqing Zong (eds.), *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 4506–4520, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.398. URL https://aclanthology.org/2020.coling-main.398.

Marzieh Fadaee, Arianna Bisazza, and Christof Monz. Data augmentation for low-resource neural machine translation. In Regina Barzilay and Min-Yen Kan (eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 567–573, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-2090. URL https://aclanthology.org/P17-2090.

Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 889–898, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1082. URL https://aclanthology.org/P18-1082.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity, 2022. URL https://arxiv.org/abs/2101.03961.

Baruch Fischhoff, Paul Slovic, and Sarah Lichtenstein. Knowing with certainty: The appropriateness of extreme confidence. *Journal of Experimental Psychology: Human perception and performance*, 3(4): 552, 1977.

Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, et al. A survey of uncertainty in deep neural networks. *ArXiv preprint*, abs/2107.03342, 2021. URL https://arxiv.org/abs/2107.03342.

Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, Muhammad Shahzad, Wen Yang, Richard Bamler, and Xiao Xiang Zhu. A survey of uncertainty in deep neural networks, 2022a.

Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, Muhammad Shahzad, Wen Yang, Richard Bamler, and Xiao Xiang Zhu. A survey of uncertainty in deep neural networks, 2022b. URL https://arxiv.org/abs/2107.03342.

Asma Ghandeharioun, Brian Eoff, Brendan Jou, and Rosalind Picard. Characterizing sources of uncertainty to proxy calibration and disambiguate annotator and data bias. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pp. 4202–4206, 2019. doi: 10.1109/ICCVW.2019.00517.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pp. 1321–1330. PMLR, 2017.

Zhen Guo, Zelin Wan, Qisheng Zhang, Xujiang Zhao, Feng Chen, Jin-Hee Cho, Qi Zhang, Lance M. Kaplan, Dong H. Jeong, and Audun Jøsang. A survey on uncertainty reasoning and quantification for decision making: Belief theory meets deep learning, 2022. URL https://arxiv.org/abs/2206.05675.

Kazuma Hashimoto, Iftekhar Naim, and Karthik Raman. How does beam search improve span-level confidence estimation in generative sequence labeling? In Raúl Vázquez, Hande Celikkanat, Dennis Ulmer, Jörg Tiedemann, Swabha Swayamdipta, Wilker Aziz, Barbara Plank, Joris Baan, and Marie-Catherine de Marneffe (eds.), *Proceedings of the 1st Workshop on Uncertainty-Aware NLP (UncertaiNLP 2024)*, pp. 62–69, St Julians, Malta, March 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.uncertainlp-1.6.

Wenchong He, Zhe Jiang, Tingsong Xiao, Zelin Xu, and Yukun Li. A survey on uncertainty quantification methods for deep learning, 2024. URL https://arxiv.org/abs/2302.13425.

Matthias Hein, Maksym Andriushchenko, and Julian Bitterwolf. Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem, 2019. URL https://arxiv.org/abs/1812.05720.

John Hewitt, Christopher Manning, and Percy Liang. Truncation sampling as language model desmoothing. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 3414–3427, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.249. URL https://aclanthology.org/2022.findings-emnlp.249.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rygGQyrFvH.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ArXiv preprint*, abs/2311.05232, 2023. URL https://arxiv.org/abs/2311.05232.

Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Machine Learning*, 110(3):457–506, March 2021. ISSN 1573-0565. doi: 10.1007/s10994-021-05946-3. URL http://dx.doi.org/10.1007/s10994-021-05946-3.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023a.

Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. Towards mitigating LLM hallucination via self reflection. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 1827–1843, Singapore, December 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.123. URL https://aclanthology.org/2023.findings-emnlp.123.

Minhao Jiang, Ken Ziyu Liu, Ming Zhong, Rylan Schaeffer, Siru Ouyang, Jiawei Han, and Sanmi Koyejo. Investigating data contamination for pre-training language models, 2024. URL https://arxiv.org/abs/2401.06059.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977, 2021. doi: 10.1162/tacl_a_00407. URL https://aclanthology.org/2021.tacl-1.57.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know, 2022.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.

Jaeyoung Kim, Kyuheon Jung, Dongbin Na, Sion Jang, Eunbin Park, and Sungchul Choi. Pseudo outlier exposure for out-of-distribution detection using pretrained transformers, 2023. URL https://arxiv.org/abs/2307.09455.

Hannah Kirk, Yennie Jun, Haider Iqbal, Elias Benussi, Filippo Volpin, Frederic A. Dreyer, Aleksandar Shtedritski, and Yuki M. Asano. Bias out-of-the-box: An empirical analysis of intersectional occupational biases in popular generative language models, 2021. URL https://arxiv.org/abs/2102.04130.

Lingkai Kong, Haoming Jiang, Yuchen Zhuang, Jie Lyu, Tuo Zhao, and Chao Zhang. Calibrated language model fine-tuning for in- and out-of-distribution data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1326–1340, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.102. URL https://aclanthology.org/2020.emnlp-main.102.

Asher Koriat, Sarah Lichtenstein, and Baruch Fischhoff. Reasons for confidence. *Journal of Experimental Psychology: Human learning and memory*, 6(2):107, 1980.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation, 2023.

Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. Diachronic word embeddings and semantic shifts: a survey. In Emily M. Bender, Leon Derczynski, and Pierre Isabelle (eds.), *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 1384–1397, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics. URL https://aclanthology.org/C18-1117.

Salem Lahlou, Moksh Jain, Hadi Nekoei, Victor Ion Butoi, Paul Bertin, Jarrid Rector-Brooks, Maksym Korablyov, and Yoshua Bengio. Deup: Direct epistemic uncertainty prediction, 2023.

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

Roger Levy. A noisy-channel model of human sentence comprehension under uncertain input. In Mirella Lapata and Hwee Tou Ng (eds.), *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pp. 234–243, Honolulu, Hawaii, October 2008. Association for Computational Linguistics. URL https://aclanthology.org/D08-1025.

Bohan Li, Yutai Hou, and Wanxiang Che. Data augmentation approaches in natural language processing: A survey. *AI Open*, 3:71–90, 2022. ISSN 2666-6510. doi: 10.1016/j.aiopen.2022.03.001. URL http://dx.doi.org/10.1016/j.aiopen.2022.03.001.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. Halueval: A large-scale hallucination evaluation benchmark for large language models, 2023a.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *ArXiv preprint*, abs/2306.03341, 2023b. URL https://arxiv.org/abs/2306.03341.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words. *ArXiv preprint*, abs/2205.14334, 2022a. URL https://arxiv.org/abs/2205.14334.

Stephanie C. Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words. *Trans. Mach. Learn. Res.*, 2022, 2022b.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Generating with confidence: Uncertainty quantification for black-box large language models. *arXiv preprint arXiv:2305.19187*, 2023.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Generating with confidence: Uncertainty quantification for black-box large language models, 2024a.

Zihao Lin, Mohammad Beigi, Hongxuan Li, Yufan Zhou, Yuxiang Zhang, Qifan Wang, Wenpeng Yin, and Lifu Huang. Navigating the dual facets: A comprehensive evaluation of sequential memory editing in large language models, 2024b. URL https://arxiv.org/abs/2402.11122.

Bo Liu, Liming Zhan, Zexin Lu, Yujie Feng, Lei Xue, and Xiao-Ming Wu. How good are llms at out-of-distribution detection?, 2024. URL https://arxiv.org/abs/2308.10261.

Tianyi Liu, Xinsong Zhang, Wanhao Zhou, and Weijia Jia. Neural relation extraction via inner-sentence noise reduction and transfer learning. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2195–2204, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1243. URL https://aclanthology.org/D18-1243.

Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks. *Advances in neural information processing systems*, 31, 2018.

Andrey Malinin and Mark Gales. Reverse kl-divergence training of prior networks: Improved uncertainty and adversarial robustness. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dÁlché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, pp. 14547–14558, 2019.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models, 2023.

Clara Meister, Ryan Cotterell, and Tim Vieira. If beam search is the answer, what was the question? In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2173–2185, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.170. URL https://aclanthology.org/2020.emnlp-main.170.

José Mena, Oriol Pujol, and Jordi Vitrià. Uncertainty-based rejection wrappers for black-box classifiers. *IEEE Access*, 8:101721–101746, 2020.

José Mena, Oriol Pujol, and Jordi Vitrià. A survey on uncertainty estimation in deep learning classification systems from a bayesian perspective. *ACM Computing Surveys (CSUR)*, 54(9):1–35, 2021.

Sabrina J. Mielke, Arthur Szlam, Emily Dinan, and Y-Lan Boureau. Reducing conversational agents' overconfidence through linguistic calibration. *Transactions of the Association for Computational Linguistics*, 10:857–872, 2022a. doi: 10.1162/tacl_a_00494. URL https://aclanthology.org/2022.tacl-1.50.

Sabrina J. Mielke, Arthur Szlam, Emily Dinan, and Y-Lan Boureau. Reducing conversational agents' overconfidence through linguistic calibration. *Transactions of the Association for Computational Linguistics*, 10:857–872, 2022b. doi: 10.1162/tacl_a_00494. URL https://aclanthology.org/2022.tacl-1.50.

Jishnu Mukhoti, Viveka Kulharia, Amartya Sanyal, Stuart Golodetz, Philip H. S. Torr, and Puneet K. Dokania. Calibrating deep neural networks using focal loss. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/aeb7b30ef1d024a76f21a1d40e30c302-Abstract.html.

Jay Nandy, Wynne Hsu, and Mong Li Lee. Towards maximizing the representation gap between in-domain & out-of-distribution examples, 2021. URL https://arxiv.org/abs/2010.10474.

Jeremy Nixon, Michael W. Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. Measuring calibration in deep learning. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 38–41. Computer Vision Foundation / IEEE, 2019. URL http://openaccess.thecvf.com/content_CVPRW_2019/html/Uncertainty_and_Robustness_in_Deep_Visual_Learning/Nixon_Measuring_Calibration_in_Deep_Learning_CVPRW_2019_paper.html.

Myle Ott, Michael Auli, David Grangier, and Marc'Aurelio Ranzato. Analyzing uncertainty in neural machine translation, 2018. URL https://arxiv.org/abs/1803.00047.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

Chanwoo Park, Mingyang Liu, Kaiqing Zhang, and Asuman Ozdaglar. Principled rlhf from heterogeneous feedback via personalization and preference aggregation. *arXiv preprint arXiv:2405.00254*, 2024.

Elise Payzan-LeNestour and Peter Bossaerts. Risk, unexpected uncertainty, and estimation uncertainty: Bayesian learning in unstable settings. *PLoS computational biology*, 7(1):e1001048, 2011.

Steven T. Piantadosi, Harry J. Tily, and Edward Gibson. The communicative function of ambiguity in language. *Cognition*, 122:280–291, 2011. URL https://api.semanticscholar.org/CorpusID:13726095.

Gwenyth Portillo Wightman, Alexandra Delucia, and Mark Dredze. Strength in numbers: Estimating confidence of large language models by prompt agreement. In Anaelia Ovalle, Kai-Wei Chang, Ninareh Mehrabi, Yada Pruksachatkun, Aram Galystan, Jwala Dhamala, Apurv Verma, Trista Cao, Anoop Kumar, and Rahul Gupta (eds.), *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pp. 326–362, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.trustnlp-1.28. URL https://aclanthology.org/2023.trustnlp-1.28.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024.

Rahul Rahaman and Alexandre H Thiery. Uncertainty quantification and deep ensembles. *stat*, 1050:20, 2020.

Matthew Renze and Erhan Guven. The effect of sampling temperature on problem solving in large language models, 2024. URL https://arxiv.org/abs/2402.05201.

Freda Shi, Daniel Fried, Marjan Ghazvininejad, Luke Zettlemoyer, and Sida I. Wang. Natural language to code translation with execution, 2022.

Chenglei Si, Chen Zhao, Sewon Min, and Jordan Boyd-Graber. Re-examining calibration: The case of question answering, 2022.

Anand Siththaranjan, Cassidy Laidlaw, and Dylan Hadfield-Menell. Distributional preference learning: Understanding and accounting for hidden context in rlhf. *arXiv preprint arXiv:2312.08358*, 2023.

Felix Stahlberg and Bill Byrne. On NMT search errors and model errors: Cat got your tongue? In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3356–3362, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1331. URL https://aclanthology.org/D19-1331.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. *arXiv preprint arXiv:2305.14975*, 2023.

Artem Vazhentsev, Akim Tsvigun, Roman Vashurin, Sergey Petrakov, Daniil Vasilev, Maxim Panov, Alexander Panchenko, and Artem Shelmanov. Efficient out-of-domain detection for sequence to sequence models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 1430–1454, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.93. URL https://aclanthology.org/2023.findings-acl.93.

Bin Wang, Chengwei Wei, Zhengyuan Liu, Geyu Lin, and Nancy F. Chen. Resilience of large language models for noisy instructions, 2024a. URL https://arxiv.org/abs/2404.09754.

Binghai Wang, Rui Zheng, Lu Chen, Yan Liu, Shihan Dou, Caishuang Huang, Wei Shen, Senjie Jin, Enyu Zhou, Chenyu Shi, Songyang Gao, Nuo Xu, Yuhao Zhou, Xiaoran Fan, Zhiheng Xi, Jun Zhao, Xiao Wang, Tao Ji, Hang Yan, Lixing Shen, Zhan Chen, Tao Gui, Qi Zhang, Xipeng Qiu, Xuanjing Huang, Zuxuan Wu, and Yu-Gang Jiang. Secrets of rlhf in large language models part ii: Reward modeling, 2024b. URL https://arxiv.org/abs/2401.06080.

Cheng Wang. Calibration in deep learning: A survey of the state-of-the-art. *ArXiv preprint*, abs/2308.01222, 2023. URL https://arxiv.org/abs/2308.01222.

Guotai Wang, Wenqi Li, Michael Aertsen, Jan Deprest, Sébastien Ourselin, and Tom Vercauteren. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing*, 338:34–45, 2019.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Jirong Wen. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6), March 2024c. ISSN 2095-2236. doi: 10.1007/s11704-024-40231-1. URL http://dx.doi.org/10.1007/s11704-024-40231-1.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models, 2023.

Junkang Wu, Yuexiang Xie, Zhengyi Yang, Jiancan Wu, Jiawei Chen, Jinyang Gao, Bolin Ding, Xiang Wang, and Xiangnan He. Towards robust alignment of language models: Distributionally robustifying direct preference optimization. *arXiv preprint arXiv:2407.07880*, 2024.

Yijun Xiao and William Yang Wang. Quantifying uncertainties in natural language processing tasks, 2018.

Yuxi Xie, Kenji Kawaguchi, Yiran Zhao, Xu Zhao, Min-Yen Kan, Junxian He, and Qizhe Xie. Self-evaluation guided beam search for reasoning, 2023. URL https://arxiv.org/abs/2305.00633.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *ArXiv preprint*, abs/2306.13063, 2023. URL https://arxiv.org/abs/2306.13063.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms, 2024.

Ruiyao Xu and Kaize Ding. Large language models for anomaly and out-of-distribution detection: A survey, 2024. URL https://arxiv.org/abs/2409.01980.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. Do large language models know what they don't know? *arXiv preprint arXiv:2305.18153*, 2023.

Hongwei Zeng, Zhuo Zhi, Jun Liu, and Bifan Wei. Improving paragraph-level question generation with extended answer network and uncertainty-aware beam search. *Information Sciences*, 571:50–64, 2021.

Ruoyu Zhang, Yanzeng Li, Yongliang Ma, Ming Zhou, and Lei Zou. LLMaAA: Making large language models as active annotators. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 13088–13103, Singapore, December 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.872. URL https://aclanthology.org/2023.findings-emnlp.872.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. Siren's song in the ai ocean: A survey on hallucination in large language models. *ArXiv preprint*, abs/2309.01219, 2023b. URL https://arxiv.org/abs/2309.01219.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.

Kaitlyn Zhou, Dan Jurafsky, and Tatsunori Hashimoto. Navigating the grey area: Expressions of overconfidence and uncertainty in language models, 2023.

Kaitlyn Zhou, Jena D. Hwang, Xiang Ren, and Maarten Sap. Relying on the unreliable: The impact of language models' reluctance to express uncertainty, 2024. URL https://arxiv.org/abs/2401.06730.

A APPENDIX

You may include other additional sections here.