Survey on Uncertainty Self-Awareness in LLMs: A Review of Recent Advances

Farbod Azimmohseni* *far.azimmohseni.82@sharif.edu*

Amirhossein Tighkhorshid* *amir.tighkhorshid78@sharif.edu*

Mahdi Mastani* *Mahdi.mastani@sharif.edu*

Danial Parnian *Danielparnian@gmail.com*

Ali Najar *anajar13750@gmail.com*

Shaygan Adim *sh83adim@gmail.com*

Department of Computer Engineering, Sharif University of Technology

Reviewed on OpenReview: . . .

Abstract

Large Language Models (LLMs) sometimes generate incorrect or hallucinated responses with unjustified confidence, making uncertainty self-awareness an important capability for reliable deployment. This survey reviews recent work on how LLMs estimate, express, and use uncertainty. We first provide the necessary background to introduce the main concepts and establish a common understanding. We then introduce the main datasets and benchmarks used in the field and discuss the advantages and limitations of each. Following this, we provide a comprehensive review of existing work across three main areas: empirical studies of uncertainty, methods for training and aligning models to better express or act upon uncertainty, and downstream applications that leverage uncertainty for tasks such as hallucination detection, abstention, adaptive reasoning, and model routing. We then compare and discuss the challenges identified across these lines of work, along with the approaches proposed to address them. Finally, in the Limitations section, we highlight the remaining limitations of the field and outline promising directions for future research.

1 Introduction

Large language models are no longer confined to research settings. They are used to summarise clinical notes, write production code, and answer questions whose answers people act on. In these settings, being right on average is not enough. A system that is right ninety percent of the time and gives no indication of which ten percent is wrong is harder to deploy safely than a weaker system that reliably flags its own doubtful cases (Kim et al., 2025). If a model is asked whether a drug interaction is dangerous, what matters is not only whether it answers correctly but whether it can tell the difference between an answer it has grounds for and one it has merely produced.

This capability is what we refer to as uncertainty self-awareness, the ability of a model to assess how likely its own output is to be correct and to communicate that assessment. We use the term in a strictly functional sense and intend no claim about consciousness or human-like introspection. It is nonetheless distinct from accuracy, and the distinction is empirical rather than philosophical. Recent work shows that a model can produce a correct answer while badly misrepresenting how sure it was, and can produce an incorrect one in language that signals complete confidence (Yona et al., 2024; Xiong et al., 2024). Hallucination and overconfidence have accompanied language models since their inception, and the most capable current models still exhibit both (Huang et al., 2025; Tonmoy et al., 2024; Rawte et al., 2023; Ji et al., 2024).

*Equal contribution.

One might expect the problem to be solved by importing the uncertainty quantification (UQ) machinery of classical machine learning, but the transfer is not straightforward for two reasons and the first reason is cost. Monte Carlo dropout and deep ensembles require repeated stochastic passes or several independently trained models, which is prohibitive at the scale of modern LLMs (Liu et al., 2025). The second is structural. In a standard classifier the output space is a fixed set of labels over which a probability distribution can be normalised, whereas in generation the space of possible outputs grows exponentially with sequence length, so computing a probability over all candidate responses is intractable (Geng et al., 2024). Worse, the token-level probabilities that are available are not measuring the right object: "Paris", "It's Paris", and "The capital of France is Paris" are three different sequences expressing one answer, so lexical diversity is counted as disagreement. Uncertainty in language models must therefore be estimated at the level of meaning rather than tokens, which is the observation that motivates much of the literature reviewed here (Kuhn et al., 2023; Farquhar et al., 2024).

Research responding to these difficulties has developed along three fairly distinct lines. One line diagnoses what current models already do, asking whether verbalized confidence tracks correctness, whether an internal signal exists that the verbal one fails to reflect, and how users respond to models' linguistic expressions of uncertainty (Xiong et al., 2024; Ni et al., 2024; Yona et al., 2024; Zhou et al., 2024). A second line intervenes in training, teaching models to state calibrated confidence, to hedge, or to abstain, on the premise that the failure is one of expression rather than of representation (Lin et al., 2022a; Zhang et al., 2024b; Xu et al., 2024; Stengel-Eskin et al., 2024). A third line treats the uncertainty estimate as an input to a decision rather than as a result, using it to detect hallucinations, to route queries between models, or to decide how much computation a question deserves (Farquhar et al., 2024; Chuang et al., 2025; Ren et al., 2025). These lines are usually surveyed separately, and one of our aims is to show that progress in one does not transfer automatically to the others.

Several surveys already cover parts of this literature. Huang et al. (2024) and Liu et al. (2025) primarily organize existing methods by estimation technique and the type of uncertainty being modeled, while Geng et al. (2024) focuses on confidence estimation and calibration. Our taxonomy instead groups works by their research objective: diagnosing whether a usable uncertainty signal exists, training models to express that signal more faithfully, or using it to support downstream decisions. This distinction is important because methods that appear directly comparable often optimize different properties, such as calibration and discrimination, or operate under different assumptions about model access. Organizing the literature by objective therefore makes these differences explicit and provides a clearer basis for comparing results across studies.

Concretely, this survey makes four contributions. First, we provide a self-contained account of the terminology, uncertainty representations, and evaluation metrics needed to interpret this literature, emphasizing the limitations that motivate each successive metric. Second, we review the datasets used across the field and identify properties that constrain current conclusions, particularly the widespread assumption that each question admits a single correct answer. Third, we organize the literature under a three-part taxonomy and compare the reviewed works across methodological choices, model-access requirements, inference cost, and training cost. Finally, we synthesize the main points of agreement and disagreement across these lines of work and identify open directions arising from the loose connection between uncertainty estimation, uncertainty expression, and downstream use.

Research on uncertainty and self-awareness in LLMs has grown rapidly in recent years, producing a literature that spans confidence estimation, calibration, uncertainty-aware training, abstention, hallucination detection, and uncertainty-guided decision making. To obtain a broad view of this literature, we constructed a bibliography through keyword-based searches covering terms related to *uncertainty*, *confidence*, *calibration*, *self-knowledge*, *abstention*, and *hallucination* in large language models, followed by screening for works directly concerned with estimating, expressing, training, or using a model's uncertainty about its own outputs. Figure **??** summarizes the resulting body of work, showing how the literature has evolved over time in terms of recurring research topics, publication venues, and publication status.

The remainder of the paper is organised as follows. Section 2 establishes the necessary background. Section 3 surveys the datasets and benchmarks. Section 4 describes how the literature was collected and presents our

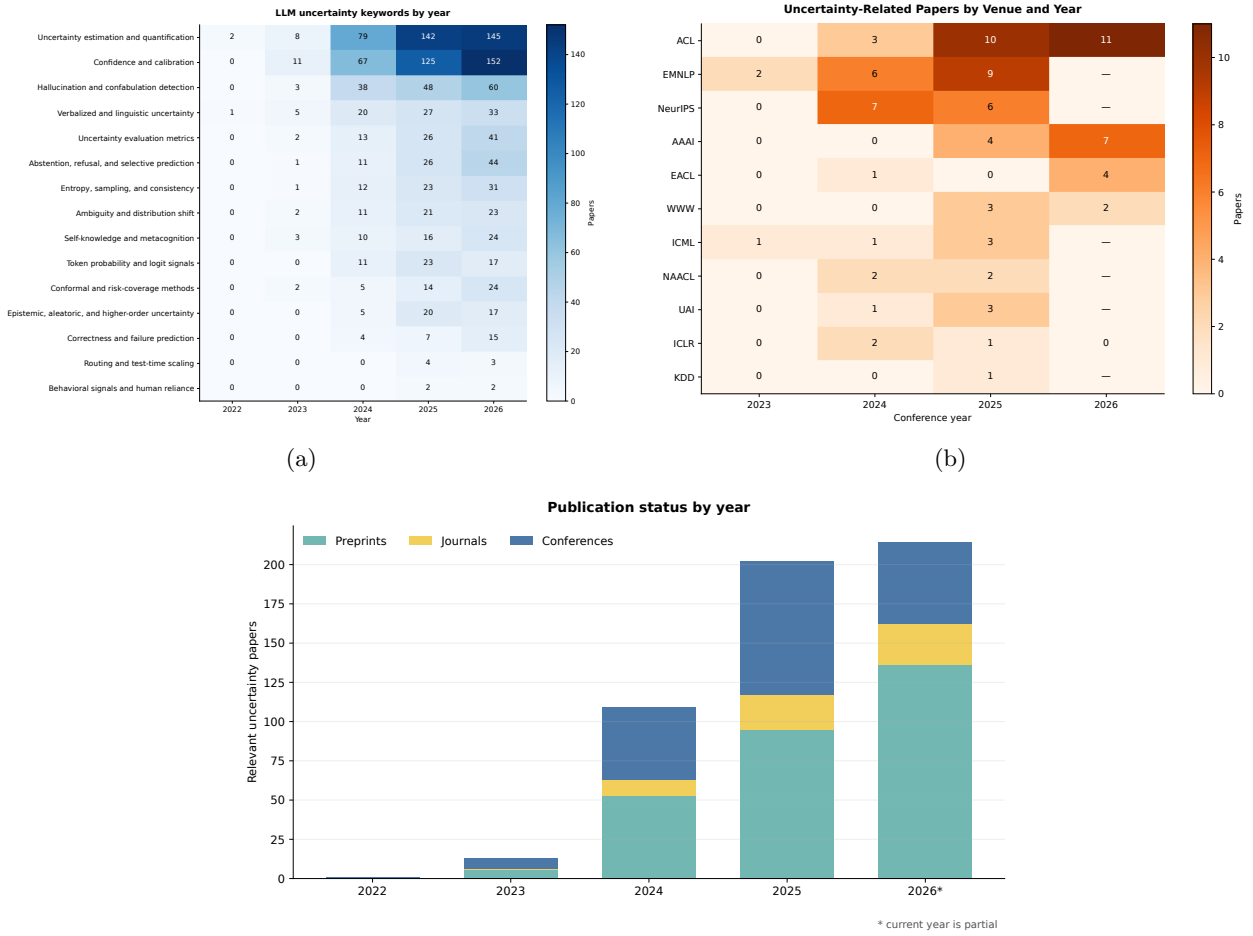

Figure 1: Overview of the collected literature: (a) uncertainty-related keywords by year, (b) uncertainty-related papers by venue and year, and (c) publication status by year.

taxonomy. Section 5 reviews the works themselves, Section 6 compares them, and Section 7 discusses open problems and future directions.

2 Background

Before reviewing the methods themselves, we establish the vocabulary that the rest of this survey uses. We provide a unified overview of the types of uncertainty, how they are conceptually represented, the methods used to elicit them from language models, and the primary metrics used for evaluation. A comprehensive and formal treatment of these concepts, including detailed mathematical formulations, and extended metric equations, is provided in Appendix A.

2.1 Aleatoric and Epistemic Uncertainty

Uncertainty in machine learning is traditionally divided into two distinct categories, and distinguishing between them matters significantly for language models because they call for fundamentally different downstream actions.

Aleatoric uncertainty arises from irreducible ambiguity, noise, or legitimate multiplicity in the data itself. Even a perfect model trained on infinite data cannot remove it. In natural language, this manifests

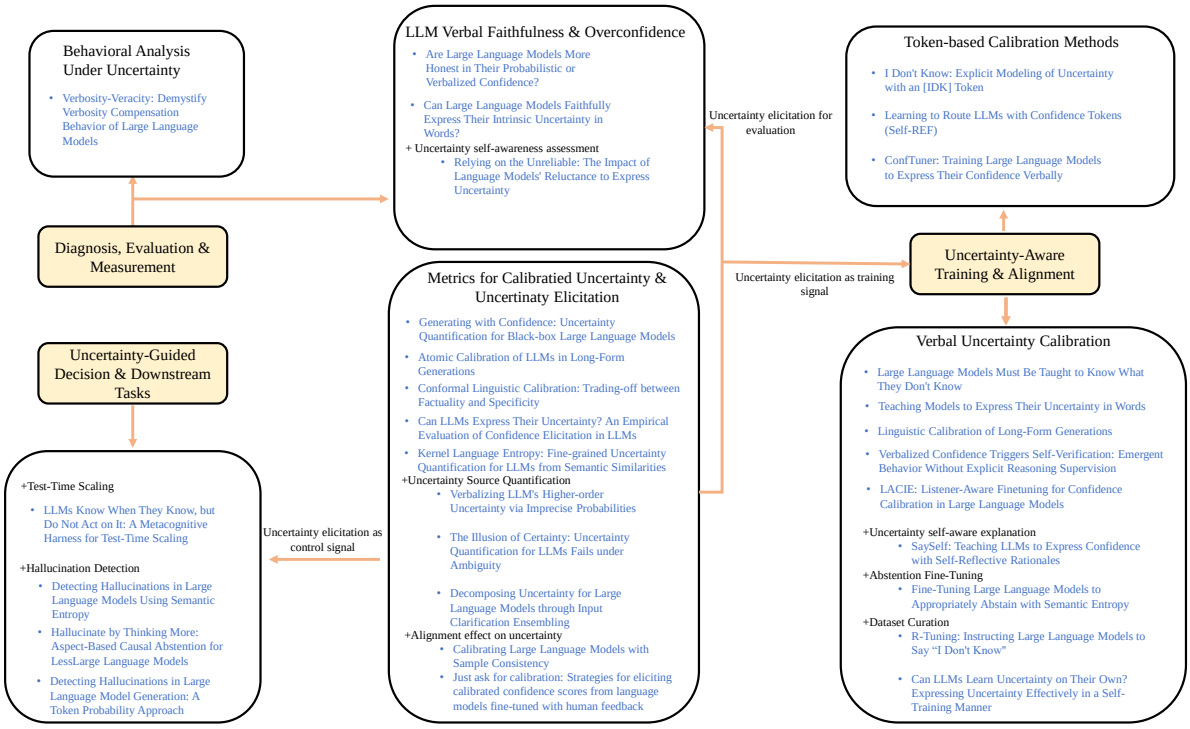

Figure 2: The given taxonomy of works in the field of uncertainty in LLMs

in questions that legitimately admit several correct answers, open-ended generative tasks where multiple diverse outputs are equally valid, or prompts that are inherently underspecified (e.g., "Who is the president of this country?").

Epistemic uncertainty, conversely, stems from the limits of the model's own knowledge. It reflects what the model does not know due to insufficient training data or queries about events occurring after its knowledge cutoff. Unlike aleatoric uncertainty, epistemic uncertainty can theoretically be reduced through additional training, better data, or external retrieval.

Most historical work on LLM self-awareness targets epistemic uncertainty to encourage models to abstain or hedge instead of fabricating answers. However, recent findings highlight that estimating aleatoric uncertainty is equally critical. Ignorance and input ambiguity both produce high disagreement among sampled answers, yet they require divergent interventions: an ignorant model should retrieve external documents or abstain, whereas an ambiguous question should be sent back to the user for clarification or answered with a broader, less specific claim that the model can confidently support (Tomov et al., 2026; Hou et al., 2024; Jiang et al., 2025).

2.2 Uncertainty Representations and Calibration

The literature currently lacks a unified definition of "confidence" and "uncertainty," generally splitting into two practical conventions.

The most common approach treats them as complementary views of a single quantity, defined as $U = 1 - C$. Under this convention, a single scalar is expected to answer two questions simultaneously: how inherently difficult or ill-posed the prompt is, and how likely the model's specific generated answer is to be correct.

A second, more expressive convention deliberately separates the two. Uncertainty is treated as an input-dependent property reflecting the query's ambiguity or difficulty. Confidence, on the other hand, is an output-dependent property describing the model's belief in one particular generated response. Conceptually, if we model sampled responses as a semantic graph, the overall connectivity of the graph indicates the

query's uncertainty, while the degree of a specific node (response) indicates the confidence in that particular answer (Lin et al., 2024).

Regardless of the adopted convention, the core objective is **calibration**, ensuring the model's stated confidence strictly aligns with its true empirical probability of being correct. For example, among all the answers a calibrated model reports with 70% confidence, exactly seventy percent should be correct. The dominant failure mode in modern LLMs is overconfidence. This overconfidence is frequently exacerbated by alignment techniques such as RLHF, which actively penalize hedged language in preference data and encourage models to sound authoritative even when incorrect (Zhou et al., 2024).

2.3 How Uncertainty Is Elicited from a Model

To utilize uncertainty, it must first be extracted. Elicitation methods fall into three modalities based on the level of model access assumed:

- **Logit-Based (Intrinsic) Uncertainty:** These methods compute statistics (like Shannon entropy or minimum token probability) directly from the predictive distribution during decoding. While computationally cheap, requiring only a single forward pass, they suffer from a major structural flaw in language generation: lexical diversity is counted as disagreement. Because "Paris" and "It's Paris" have different token sequences, token-level entropy may appear high even when the underlying meaning is certain.
- **Sampling-Based Uncertainty:** To capture meaning-level variability, these methods generate multiple responses under stochastic decoding and measure semantic agreement (Lyu et al., 2025). Approaches like semantic entropy group responses into equivalent meaning clusters using an auxiliary Natural Language Inference (NLI) model, making the estimate invariant to paraphrase (Kuhn et al., 2023; Farquhar et al., 2024). While highly informative, these methods are computationally expensive, requiring k generations and $O(k^2)$ semantic comparisons (Lin et al., 2024).
- **Verbalized (Linguistic) Confidence:** For proprietary, black-box APIs, models can be prompted to express their confidence linguistically. This can take the form of numerical scoring, explicit hedging ("I am not entirely sure, but..."), or post-hoc self-probing where the model evaluates its own previous answer. While universally deployable, empirical studies consistently find verbalized estimates to be poorly calibrated and highly sensitive to prompt wording (Xiong et al., 2024; Ni et al., 2024).

2.4 Metrics for Evaluating Uncertainty

Evaluating an uncertainty estimator requires metrics that capture distinct facets of reliability:

- **Expected Calibration Error (ECE):** The oldest metric, ECE partitions predictions into confidence bins and measures the average gap between stated confidence and empirical accuracy within each bin (Naeini et al., 2015; Guo et al., 2017).
- **AUROC and Discrimination:** A model can achieve a perfect ECE by assigning the exact same confidence score to every sample if that score matches its average accuracy, rendering the estimate useless for identifying specific errors (Xiong et al., 2024). The Area Under the ROC Curve (AUROC) addresses this by measuring *discrimination*,the probability that the estimator assigns higher uncertainty to a randomly chosen incorrect prediction than to a correct one.
- **Risk-Coverage and Abstention:** When a model is allowed to abstain, the trade-off between how much it answers (coverage) and how often it is wrong (risk) becomes central (Geifman & El-Yaniv, 2017). Metrics like the Accuracy-Engagement Distance (AED) evaluate this balance, rewarding models that answer accurately while successfully declining cases they cannot handle (Tjandra et al., 2024).

Finally, a recent line of work argues for **higher-order uncertainty**, separating uncertainty over the possible answers from uncertainty about the distribution itself (Yang et al., 2026; Tomov et al., 2026). Traditional estimators typically report first-order uncertainty, which presupposes that the model has a single, well-determined probability distribution over its candidate answers. Second-order uncertainty, however, arises when the model is so fundamentally uncertain that it cannot even settle on how likely its own answers are. By representing confidence as a probability interval rather than a scalar point estimate, the width of this interval explicitly measures how indeterminate the model's underlying belief is. This richer representation allows models to distinguish between two fundamentally different situations that a single scalar collapses into one: confidently believing a specific answer is probably wrong, and having absolutely no empirical basis for assigning a probability to it at all.

3 Datasets and Benchmarks

Almost every claim in this literature relies on benchmarks originally built for question answering or reasoning, later repurposed to provide ground-truth answers for scoring confidence estimates. This section categorises the primary datasets driving uncertainty research and highlights their inherent limitations. A comprehensive, dataset-by-dataset breakdown of all benchmarks reviewed—including their specific use cases, advantages, and configurations—is provided in Appendix B, while Table 3 provides a high-level summary.

3.1 Knowledge, Reasoning, and Domain-Specific Benchmarks

The core of the field relies on closed-book factual QA and reading comprehension to test parametric knowledge. Benchmarks like **TriviaQA** (Joshi et al., 2017) and Natural Questions (**NQ**) (Kwiatkowski et al., 2019) are heavily utilized for evaluating semantic entropy and comparing probabilistic versus verbalized confidence (Farquhar et al., 2024; Lin et al., 2024; Ni et al., 2024; Yona et al., 2024). To test whether uncertainty estimates degrade under domain shift, expert-curated sets like **BioASQ** (Krithara et al., 2023) and **MedQA** (Jin et al., 2021) are commonly employed (Chuang et al., 2025; Band et al., 2024).

Additionally, because uncertainty arises from both factual ignorance and unstable derivation steps, reasoning tasks such as **GSM8K** (Cobbe et al., 2021) and **HotpotQA** (Yang et al., 2018) serve to isolate confidence in multi-step problem solving (Xiong et al., 2024; Lyu et al., 2025; Xu et al., 2024). Multiple-choice suites like **MMLU** (Hendrycks et al., 2021) provide convenient token-level confidence targets, though they artificially inflate apparent calibration by enforcing a baseline accuracy floor and preventing abstention (Zhang et al., 2024b; Cohen et al., 2024).

3.2 Targeted Failure Modes, Ambiguity, and Long-Form Generation

A separate family of datasets is constructed to expose specific vulnerabilities. **TruthfulQA** (Lin et al., 2022b), **HaluEval** (Li et al., 2023), and **SelfAware** (Yin et al., 2023) target situations where models are prone to hallucinate, mimic misconceptions, or fail to recognize unanswerable queries (Stengel-Eskin et al., 2024; Zhang et al., 2024b).

Crucially, most standard benchmarks assume a single correct answer, conflating genuine input ambiguity with model ignorance. Datasets like **AmbigQA** (Min et al., 2020) and **MAQA*** (Tomov et al., 2026) break this assumption, demonstrating that standard uncertainty proxies often fail when a prompt permits multiple valid interpretations (Hou et al., 2024; Tomov et al., 2026). Finally, to combat the fact that short-answer scoring overstates reliability, long-form benchmarks such as FactScore biographies (Min et al., 2023) and **Qasper** (Dasigi et al., 2021) are increasingly used to evaluate atomic-claim calibration and verbosity compensation (Zhang et al., 2024a;c; Tjandra et al., 2024).

3.3 Limitations of the Current Dataset Landscape

Four properties currently constrain the field's conclusions:

- **Format Skew:** An over-reliance on short-form English factoid QA, which does not reflect realistic, multilingual, or long-form deployments.

- **Single-Answer Assumption:** The systemic presupposition of a single correct answer, which prevents estimators from distinguishing between prompt ambiguity and model ignorance (Tomov et al., 2026).

- **Data Contamination:** Wikipedia-derived benchmarks are near-certain to appear in the pretraining corpora of evaluated models, artificially shrinking the low-confidence regions methods aim to detect.

- **Evaluation Bottlenecks:** The increasing reliance on LLM-as-a-judge protocols (e.g., in long-form settings or semantic entropy evaluations) introduces a second, uncalibrated source of error into the very ground-truth labels used to score calibration.

4 Scope and Organisation of the Survey

Existing surveys typically organise this area by technique, separating white-box from black-box estimators, or single-pass from sampling-based ones Liu et al. (2025); Geng et al. (2024). Such taxonomy might obscures the field's actual disagreements, as two papers using the same technique may be answering entirely different questions and two papers answering the same question may reach opposite conclusions simply because one reports calibration error and the other reports discrimination. We therefore organise the literature by what each work is trying to achieve, into the three groups shown in Figure 2.

1. **Empirical findings and diagnostic evaluation.** Work in this group asks whether current models already possess an internal representation of their own uncertainty, and whether this internal uncertainty is reliably reflected in their observable outputs. It measures the mismatch between internal probability distributions and verbalized statements across datasets, question frequencies, and prompting strategies, and increasingly asks whether the estimators themselves are measuring what they claim to measure. Here, uncertainty is the object of measurement.

2. **Training and alignment for uncertainty elicitation.** Work in this group assumes that models contain a usable internal uncertainty signal but do not necessarily express it in a calibrated or interpretable form. These approaches use supervised fine-tuning, reinforcement learning, preference optimization, or specialized output representations to align model responses with their underlying uncertainty. The goal is to produce calibrated confidence estimates, appropriate expressions of uncertainty, explanations of uncertainty sources, or explicit abstention when the model lacks sufficient knowledge. Here, uncertainty is the training target.

3. **Downstream applications of uncertainty signals.** Work in this group assumes that an uncertainty estimate is available and investigates how it can inform downstream decisions. These approaches use uncertainty to identify potentially hallucinated responses, determine when a model should abstain, allocate test-time computation, or route queries among models with different capabilities and costs. Here, uncertainty serves as an input to a decision policy, and performance is evaluated by the quality of the resulting decisions rather than by the accuracy or calibration of the uncertainty estimate itself.

The three groups are ordered as a rough historical trajectory, but they are not a pipeline, and one of our central observations is that they remain only loosely coupled. An estimator can be excellent and say nothing about how to phrase the result; a model can be trained to hedge convincingly without its underlying estimate improving; and a router can work well with a signal whose absolute calibration is poor. Section 6 develops this point, and Section 7 argues that the most valuable open problems lie precisely at the boundaries between the three.

5 Review of Key Papers

This section reviews the most influential works on uncertainty awareness and self-knowledge in Large Language Models (LLMs). Following the taxonomy introduced in Section 4, we organize the literature into three categories: (1) empirical investigations of uncertainty expression, (2) training and alignment methods for uncertainty calibration, and (3) downstream applications that exploit uncertainty signals.

5.1 Empirical Findings and Diagnostic Evaluation

A persistent limitation of LLMs is their tendency to express incorrect answers with unwarranted confidence. As these models became increasingly capable and widely used, this behavior attracted growing attention because model errors are considerably more difficult to manage when the model provides little indication that its answer may be unreliable. This motivated a growing body of work on whether LLMs can recognize the limits of their own knowledge and distinguish reliable answers from those that are likely to be incorrect. Empirical studies have repeatedly found that models tend to produce overly confident or decisive responses even when their answers are incorrect or when other signals indicate substantial uncertainty (Xiong et al., 2024; Yona et al., 2024; Zhou et al., 2024; Ni et al., 2024). Consequently, research in this area has focused increasingly on characterizing the relationship between model uncertainty, expressed confidence, and actual correctness.

Initial empirical studies examined whether confidence could be elicited directly from the model. Tian et al. (2023) showed that models fine-tuned with human feedback can produce meaningfully calibrated verbal confidence under appropriate elicitation strategies, with calibration improving under prompting procedures that encourage additional reasoning or comparison between alternatives. Xiong et al. (2024) provided a broader evaluation across prompting, sampling, and aggregation strategies and found that verbalized confidence remains strongly biased toward high values. More importantly, they distinguished calibration from failure prediction. A model can exhibit reasonable aggregate calibration while still being unable to assign lower confidence specifically to its incorrect answers. This distinction established that evaluating uncertainty awareness requires considering not only whether reported confidence matches average accuracy, but also whether it can discriminate between reliable and unreliable predictions.

An early direction therefore focused on whether models could directly communicate how confident they were in their answers. Ni et al. (2024) compared two forms of confidence: verbalized confidence, in which the model explicitly states how certain it is, and probabilistic confidence derived from its predictive probabilities. They found that probabilistic confidence was more closely aligned with correctness, whereas verbalized confidence showed weaker and less consistent correspondence with actual performance. This comparison provided early evidence that the confidence expressed in language may not faithfully reflect the information available from the model's predictive distribution Uncertainty can also appear indirectly in the form of generation behavior rather than explicit confidence statements.

Subsequent work examined this discrepancy more directly by asking whether expressed confidence faithfully reflects the model's own uncertainty. Yona et al. (2024) compared the decisiveness of a model's language with its intrinsic uncertainty, estimated from agreement across repeated generations. They found that models frequently produce decisive responses even when repeated sampling reveals substantial disagreement, showing that expressed confidence can be poorly aligned with the model's own generation behavior. Zhou et al. (2024) further demonstrated why this mismatch matters in practice. Users interpret confident or unhedged language as evidence of reliability, including when the underlying answer is incorrect. Together, these findings show that uncertainty evaluation must consider not only whether a useful estimate can be obtained from a model, but also whether the confidence communicated to users reflects that estimate.

Zhang et al. (2024c) identify verbosity compensation, showing that models tend to produce unnecessarily long answers when they are less certain. Across multiple knowledge and reasoning tasks, more verbose responses were associated with lower accuracy and higher uncertainty, suggesting that response length itself can serve as a behavioral signal of unreliability. This finding broadens the study of uncertainty expression beyond explicit confidence scores by showing that uncertainty may be reflected in how a model answers even when it is not stated directly..

Motivated by this gap, another line of work avoids relying on explicit confidence statements and instead estimates uncertainty from the model's generation behavior. The central intuition is that repeated generations provide indirect evidence about how stable the model's prediction is. Consistent answers suggest greater confidence, whereas disagreement across samples indicates uncertainty. Lyu et al. (2025) systematically studied this approach through sample consistency, comparing several agreement-based measures across multiple models and reasoning tasks. Their results show that consistency-based confidence can outperform verbalized confidence, self-evaluation, and token-level probability baselines in identifying unreliable predictions. At the same time, this approach exposes an important limitation of behavioral uncertainty estimation. A model that repeatedly produces the same incorrect answer may appear highly confident despite being wrong.

A difficulty with sample consistency is that disagreement at the surface level does not necessarily imply disagreement in meaning. Different generations may use different wording while expressing the same underlying answer, causing lexical variation to be mistaken for uncertainty. This motivated a shift from comparing generated strings directly to comparing their semantic content. Kuhn et al. (2023) introduced semantic uncertainty by grouping sampled responses according to semantic equivalence and measuring uncertainty over these meaning-level groups rather than over individual sequences. Farquhar et al. (2024) further demonstrated the usefulness of this formulation for detecting confabulations, showing that semantic entropy can distinguish unreliable generations more effectively than uncertainty measures based purely on token probabilities or lexical variation. The key contribution of this line of work is therefore not simply the use of multiple samples, but the recognition that uncertainty in language generation should be measured over distinct meanings rather than distinct strings.

Later work refined this idea by relaxing the assumption that semantic relationships between generations must be discrete. Semantic entropy effectively treats two responses as either equivalent or different, whereas in practice semantic similarity can vary continuously. Nikitin et al. (2024) address this limitation with Kernel Language Entropy, representing sampled responses through pairwise semantic similarities and estimating uncertainty from the resulting kernel structure. In a related black-box setting, Lin et al. (2024) construct semantic similarity graphs over sampled generations and use their structure to estimate both predictive uncertainty and response-level confidence without requiring access to model logits or hidden representations.

Despite these refinements, disagreement among generations does not have a unique interpretation. A model may produce different answers because it lacks sufficient knowledge, but the same pattern can arise when the question itself is ambiguous or admits several valid answers. Hou et al. (2024) make this distinction explicit by separating uncertainty associated with the input from uncertainty attributable to the model. Their Input Clarification Ensembling approach generates alternative clarifications of an ambiguous query and examines the variation both across and within these clarified versions. Disagreement across clarifications is used to capture uncertainty arising from the question itself, whereas disagreement that remains within a clarification is treated as evidence of uncertainty in the model's prediction. This decomposition is important because the two sources of uncertainty call for different responses: ambiguity may be resolved by clarifying the user's intent, while insufficient model knowledge may instead require abstention, retrieval, or additional reasoning.

Tomov et al. (2026) push this argument further by showing that the ambiguity problem is not merely a limitation of a particular estimator, but a more general failure mode of consistency-based uncertainty quantification. They demonstrate that estimators based on agreement among sampled responses implicitly rely on the assumption that each query has a single correct answer. When multiple answers are legitimately possible, increased sample diversity no longer provides reliable evidence that the model is uncertain about its knowledge. Their experiments on ambiguity-aware benchmarks show that methods performing well on conventional single-answer question answering can degrade substantially once this assumption is relaxed. These findings place an important qualification on earlier consistency- and semantic-based results: disagreement is informative only when the source of that disagreement is understood.

The ambiguity results also expose a limitation in the common practice of representing uncertainty with a single scalar. A single confidence value may indicate that a prediction is unreliable, but it does not reveal whether this unreliability arises from ambiguity in the input, insufficient model knowledge, or uncertainty about the confidence estimate itself. Yang et al. (2026) address this limitation through imprecise probabilities, allowing models to express confidence as an interval rather than as a single point estimate. The width and

location of this interval provide additional information about the model's uncertainty and make it possible to distinguish uncertainty about an answer from uncertainty about the confidence assigned to that answer. Their results suggest that richer uncertainty representations can improve both ambiguity detection and correctness prediction while remaining compatible with black-box prompting.

A related issue is that uncertainty need not be communicated only through an explicit confidence score. It can also affect how specific a model is willing to be in its answer. Jiang et al. (2025) study this trade-off between factuality and specificity, observing that methods which improve factual reliability can do so simply by producing less specific statements. They introduce a conformal linguistic calibration framework that adjusts the specificity of a response according to its estimated uncertainty, allowing the model to make precise claims when sufficient evidence is available and broader claims when confidence is lower. This shifts the role of uncertainty from merely describing the reliability of a fixed answer to controlling the strength of the claim that the model is willing to make. It also highlights an important evaluation issue: improvements in factuality should not be interpreted independently of how much information the model retains in its response.

Most of the preceding methods assign a single uncertainty estimate to an entire response, which can hide variation in reliability across individual claims. Zhang et al. (2024a) address this limitation by evaluating calibration at the atomic-claim level in long-form generations. Their results show that responses appearing well calibrated as a whole can still contain individual statements with substantially different levels of reliability, motivating finer-grained uncertainty estimation for long-form outputs.

5.2 Training and Alignment for Uncertainty Elicitation

The persistent overconfidence observed in LLMs has motivated a growing line of work that moves beyond post-hoc uncertainty estimation and instead explicitly trains models to recognize or communicate their uncertainty. These methods target different behaviors, including calibrated confidence expression, correctness prediction, abstention, and explanations of uncertainty. This distinction is important because conventional alignment objectives primarily optimize helpful and correct responses and do not necessarily encourage models to preserve or faithfully communicate uncertainty. Indeed, some studies suggest that standard supervised or preference-based alignment can encourage models to answer authoritatively even when their underlying knowledge is insufficient (Zhang et al., 2024b; Zhou et al., 2024; Stengel-Eskin et al., 2024).

A major motivation for uncertainty-aware training is also computational. White-box uncertainty estimators require access to model probabilities or representations, while sampling-based approaches often require multiple generations and semantic comparisons. Training-based methods can shift part of this cost to an offline stage where potentially expensive uncertainty estimators are used to construct supervision, after which the fine-tuned model produces an uncertainty signal directly during generation. Existing approaches differ, however, in the source of this supervision, how the training data are constructed, the optimization objective, and the form of uncertainty ultimately produced by the model.

Early work demonstrated that confidence itself can be treated as a supervised prediction target. Lin et al. (2022a) constructed uncertainty-labeled data from empirical model accuracy and fine-tuned GPT-3 to generate numerical or verbal confidence alongside its answers. Besides improving calibration, the authors observed greater separation between hidden representations of correct and incorrect predictions, suggesting that uncertainty supervision can also affect internal representations. Kapoor et al. (Kapoor et al., 2024) similarly showed that uncertainty-related behavior can be learned from relatively small amounts of graded supervision, although their objective is more directly correctness prediction. They do LoRA fine-tuning with linguistic verification prompts produced predictors that generalized across domains and even across outputs from different model families.

Subsequent methods differ more substantially in how they transform correctness or uncertainty estimates into training examples. R-Tuning (Zhang et al., 2024b) uses the base model's own performance to partition an instruction-tuning dataset into examples that the model already answers correctly and those that it does not. The two subsets are then associated with "sure" and "unsure" responses, respectively. Its central argument is that ordinary supervised fine-tuning may blur the model's original knowledge boundary by

teaching every training example as if it were equally known. Instead, R-Tuning makes this boundary part of the supervision.

UaIT (Liu et al., 2024) uses a richer uncertainty signal rather than binary correctness alone. It applies a semantic weight to model-generated responses, constructs confidence labels from these uncertainty estimates, and filters the training set so that correct/high-confidence and incorrect/low-confidence examples are retained while inconsistent examples are removed. The resulting model is instruction-tuned to generate both its answer and confidence in one pass. Uncertainty Distillation (Hager et al., 2025) follows a related teacher-student principle but derives supervision from semantic agreement across multiple generations. The resulting estimates are calibrated and discretized into verbal confidence categories, with controlled inclusion of incorrect examples to teach low-confidence behavior without substantially degrading answer accuracy. Mostly, both approaches move expensive uncertainty estimation to dataset construction, but differ in the uncertainty estimator, calibration procedure, and filtering strategy used to produce the training targets.

The same distinction appears in methods that train uncertainty as an action rather than as a confidence value. Different from R-Tuning that obtains its uncertain subset from whether the base model answers correctly, Tjandra et al. (2024) use semantic entropy to identify high-uncertainty examples and convert them into abstention supervision. Both methods therefore teach models not to respond confidently in unreliable regions, but they define these regions differently where R-Tuning approximates the model's knowledge boundary through correctness, while semantic-entropy fine-tuning uses disagreement among semantically distinct sampled answers. The latter can provide richer uncertainty information, but incurs the additional cost of multi-sample uncertainty estimation during dataset construction.

Methods also differ in the richness of the uncertainty behavior they attempt to teach. Most approaches produce a scalar confidence value or a short certainty/uncertainty statement, whereas SaySelf (Xu et al., 2024) jointly trains models to express confidence and generate self-reflective rationales explaining the source of their uncertainty. It first uses supervised fine-tuning to learn these outputs and subsequently applies PPO to better align confidence with correctness. Uncertainty communication is extended further by Xiong et al. (2024) to long-form generation, where a single response-level confidence value may be insufficient. Their framework first distills confidence-aware summaries from multiple generations and then applies decision-based reinforcement learning so that the uncertainty expressed throughout a response supports calibrated judgments by a reader.

The choice of optimization objective provides another important axis of variation. Many methods retain ordinary supervised cross-entropy and primarily modify the labels or training data, whereas others modify the optimization procedure itself. SaySelf and long-form linguistic calibration introduce reinforcement-learning stages after supervised training, while LACIE (Stengel-Eskin et al., 2024) formulates uncertainty communication as a preference-learning problem. Rather than directly supervising a numerical confidence target, LACIE uses a simulated listener to determine how confident a response appears and constructs DPO preferences that strongly penalize confidently communicated incorrect answers. Interestingly, this listener-aware objective also induces abstention behavior without requiring explicit "I don't know" demonstrations. These approaches illustrate that calibrated uncertainty can be encouraged either by imitating uncertainty labels or by optimizing the consequences of how confidence is communicated.

Uncertainty supervision can additionally alter the model's reasoning behavior, rather than only its final confidence statement. Jang et al. (Jang et al., 2025) demonstrate this with Confidence-Supervised Fine-Tuning (CSFT), where empirical confidence labels are obtained from consistency across sampled reasoning traces and the supervised loss is applied primarily to the confidence suffix rather than the reasoning trajectory itself. Despite the absence of explicit self-verification supervision, low predicted confidence leads the model to generate longer reasoning traces and to recheck intermediate steps. CSFT therefore differs from methods such as UaIT or Uncertainty Distillation in an important aspect. Although all train confidence from automatically generated supervision, CSFT investigates how confidence learning feeds back into the generation process itself.

The methods above largely represent uncertainty through natural-language confidence statements, rationales, or abstention behavior. A related line of work instead makes uncertainty an explicit component of the model's

token space. This formulation allows the probability assigned to uncertainty tokens to be directly inspected and enables losses to target uncertainty separately from the answer sequence.

IDK-tuning (Cohen et al., 2024) introduces a dedicated `[IDK]` token and modifies the next-token training target so that probability mass is shifted toward this token when the model is likely to make an incorrect prediction. Consequently, uncertainty supervision is incorporated directly into the language-modeling objective without requiring separately annotated confidence labels. Self-REF Xu et al. (2024) instead introduces binary `<CN>` and `<UN>` tokens corresponding to correct and uncertain predictions. Its distinguishing feature is gradient masking over incorrect answer sequences where the model learns to emit the uncertainty token from its errors without simultaneously being trained to reproduce those erroneous answers. The relative probabilities of the two tokens can then be interpreted as a continuous confidence score and used for selective routing.

ConfTuner (Li et al., 2025) also represents confidence through a restricted set of confidence tokens but differs in its training objective. Rather than assigning each example a heuristic confidence label, it uses answer correctness together with a tokenized Brier-style loss that directly rewards calibrated confidence distributions. So, IDK-tuning, Self-REF, and ConfTuner all expose uncertainty through dedicated tokens, but they obtain the learning signal differently. IDK-tuning modifies the target distribution according to the model's uncertainty, Self-REF learns binary confidence tokens from error feedback while masking incorrect generations, and ConfTuner optimizes calibration directly from correctness through a proper scoring objective.

Finally, these studies show that uncertainty-aware alignment is not a single fine-tuning strategy but a collection of choices from the source of uncertainty signal, the supervision level the signal is used for training, and the target behaviour of the model after alignment. At the supervision level, methods range from binary answer correctness (Zhang et al., 2024b; Li et al., 2025) to sampling-based semantic uncertainty (Hager et al., 2025; Tjandra et al., 2024) and probability-derived estimates (Liu et al., 2024). At the behavioral level, training targets range from numerical confidence and linguistic hedging to abstention, self-reflective explanations, and explicit uncertainty tokens. Finally, while conventional supervised fine-tuning remains the most used mechanism, reinforcement learning, preference optimization, gradient masking, and calibration-specific scoring objectives demonstrate that the training objective itself can substantially affect how uncertainty is learned and expressed.

5.3 Downstream Applications of Uncertainty Signals

Once an uncertainty signal can be estimated or elicited, it can be used to improve downstream decisions rather than treated as an end in itself. Existing work primarily uses uncertainty for hallucination detection, model routing, adaptive reasoning, and abstention, where its practical value depends less on absolute calibration than on whether it supports better decisions under accuracy, latency, and computational constraints.

A direct application is hallucination detection, where uncertainty serves as a signal for identifying unreliable answers. Manakul et al. (2023) use disagreement across repeated generations, while Quevedo et al. (2024) derive lightweight probability-based features for distinguishing hallucinated from reliable responses. Semantic uncertainty provides a stronger meaning-level signal in Kuhn et al. (2023) and Farquhar et al. (2024), while Chen et al. (2024) exploit internal representations to predict hallucination risk. These approaches ultimately frame uncertainty as a ranking or classification signal for deciding which responses require verification or rejection.

Another major application is model routing and adaptive computation. Zhang et al. (2024c) and Gupta et al. (2024) use uncertainty to determine when queries should be escalated to more capable models, while Chuang et al. (2025) and Li et al. (2025) use learned confidence signals for cascading and selective self-correction. Similarly, Ren et al. (2025) use uncertainty to allocate additional reasoning only when it is likely to improve the answer, and Nguyen et al. (2026) use aspect-level uncertainty to support more informed abstention decisions. Across these applications, uncertainty therefore acts primarily as a control signal that determines whether to answer, abstain, reason further, or invoke additional computational resources.

A central challenge, however, is the cost of obtaining reliable uncertainty estimates. Expensive sampling- and semantic-based estimators can provide stronger signals but undermine the computational savings sought by

routing and adaptive inference, whereas cheaper behavioral or probabilistic signals may be less reliable. This creates a fundamental trade-off between uncertainty quality and the cost of estimating it, motivating methods that evaluate uncertainty not only by calibration but also by its effectiveness in downstream decision-making.

6 Comparison and Discussion

The reviewed literature can be understood as a progression from identifying uncertainty signals to learning how to communicate them and, finally, using them to control model behavior. Rather than treating these works as isolated approaches, we organize them according to the type of signal they exploit and the role that uncertainty plays in the system.

Finding and characterizing uncertainty. The first group of studies investigates whether models contain useful uncertainty signals and how these signals can be extracted. A straightforward line of work elicits confidence directly from the model through prompting and verbalized probabilities (Tian et al., 2023; Xiong et al., 2024). These studies show that models can communicate confidence, but the expressed confidence is often poorly aligned with actual correctness. This motivated approaches that rely less on what the model says and more on signals available from its predictions. Token probabilities provide a relatively direct white-box signal and can be obtained from a single generation (Ni et al., 2024; Quevedo et al., 2024), while repeated sampling uses disagreement across generations as a proxy for uncertainty (Lyu et al., 2025; Yona et al., 2024). The latter is particularly useful when model probabilities are unavailable, but requires multiple forward passes.

A further line of work moves from surface-level disagreement to semantic disagreement. Semantic entropy and related kernel- or graph-based methods group generations according to their meaning rather than their exact wording, allowing uncertainty to reflect whether the model produces genuinely different answers rather than merely different phrasings (Farquhar et al., 2024; Nikitin et al., 2024; Lin et al., 2024). Input-clarification methods extend this idea in a different direction: instead of repeatedly sampling the same question, they perturb or clarify the input and examine which disagreements remain after the ambiguity of the question has been reduced (Hou et al., 2024). This provides a way to distinguish uncertainty arising from the input from uncertainty arising from the model itself. Other studies further show that uncertainty can manifest through behavioral properties such as verbosity (Zhang et al., 2024c) or through the distinction between probabilistic and verbalized confidence (Ni et al., 2024; Yona et al., 2024).

Overall, these approaches reveal a clear trade-off. Simple signals such as verbalized confidence or token probabilities are cheap but can be poorly calibrated or require white-box access. Sampling-based and semantic methods provide richer uncertainty estimates and are generally more faithful to the model's predictive behavior, but their computational cost grows with the number of generations and, for semantic methods, with the comparisons required between them. This trade-off motivates the second line of research: instead of estimating an expensive uncertainty signal at inference time, can the signal be learned during training?

Training and alignment for uncertainty. The second group of methods attempts to move uncertainty estimation from inference-time measurement into the training process. Early work uses supervised fine-tuning to directly teach models to associate answers with calibrated verbal confidence (Lin et al., 2022a). Later approaches expand this idea by constructing training data from model-generated samples or probabilistic uncertainty and using distillation or reinforcement learning to teach models to express uncertainty in natural language (Band et al., 2024; Liu et al., 2024). The supervision signal therefore changes from the model's own verbalized confidence to an externally estimated measure of correctness or uncertainty.

A different strategy is to train the model around its knowledge boundary rather than explicitly predicting a confidence score. Refusal-aware tuning partitions questions according to whether their answers are supported by the model's knowledge and trains the model to abstain when appropriate (Zhang et al., 2024b). Semantic-entropy-based training follows a stronger teacher-signal approach: expensive semantic uncertainty estimation is performed offline and the resulting signal is used to train the model to make a one-pass uncertainty or abstention decision (Tjandra et al., 2024; Hager et al., 2025). This improves inference efficiency, but shifts the computational burden to dataset construction and training.

Other methods encode uncertainty directly into the model's output space. Explicit tokens such as `[IDK]` or dedicated confidence tokens provide a compact machine-readable signal that can subsequently be used for selective prediction or routing (Cohen et al., 2024; Li et al., 2025; Chuang et al., 2025). Preference-based approaches instead optimize how uncertainty is communicated to a listener, making the objective sensitive not only to correctness but also to how a human interprets the response (Stengel-Eskin et al., 2024). Similarly, self-reflective rationales and scalar confidence supervision attempt to make uncertainty expression part of the model's normal generation behavior (Xu et al., 2024; Jang et al., 2025).

These approaches therefore differ mainly in what they use as the supervision signal and where they place the computational cost. Direct confidence supervision is simple but depends on the quality of the confidence labels. Semantic-uncertainty supervision provides a richer target because it captures meaning-level disagreement, but requires expensive offline sampling and semantic comparison. Dedicated uncertainty tokens are particularly attractive when uncertainty is consumed by another machine component, such as a router or abstention policy, whereas natural-language calibration is more appropriate when uncertainty must be communicated to a human.

Downstream use of uncertainty. Once an uncertainty signal is available, a third group of studies uses it to modify system behavior. Sampling-based and semantic uncertainty estimates have been used for hallucination and confabulation detection (Manakul et al., 2023; Farquhar et al., 2024), while token probabilities provide a cheaper alternative for detecting unreliable generations (Quevedo et al., 2024). Other approaches use uncertainty to decide whether a model should answer or defer to another model, enabling cascade and routing strategies (Gupta et al., 2024). More recent work extends this idea to adaptive reasoning and aspect-based abstention, where uncertainty determines whether the model should spend additional computation or refuse a particular claim (Ren et al., 2025; Nguyen et al., 2026).

This progression highlights that the value of uncertainty ultimately depends on the decision it enables. A highly accurate uncertainty estimator is not necessarily useful if obtaining it costs more than the computation it is intended to save. Conversely, a lightweight confidence signal can be valuable for routing or filtering even if it is less expressive, provided that it preserves the ordering of reliable and unreliable predictions. The appropriate uncertainty representation should therefore be evaluated jointly with its computational requirements and downstream decision.

To consolidate the findings of our review, Table **??** provides a structured comparison of the reviewed works across their primary focus and methodological characteristics, highlighting how different approaches address uncertainty estimation, faithfulness, behavioral analysis, abstention, semantic modeling, and sampling. Complementing this qualitative comparison, Table **??** summarizes the practical requirements of the main method families in terms of model access, inference cost, training cost, and representative works. Together, these tables provide a unified view of both the methodological landscape and the practical trade-offs underlying existing uncertainty-awareness approaches.

Table 1: Comparison of the reviewed uncertainty-awareness works. Column definitions are given at the end of the table.

Work	Primary focus	FT	DC	FA	BA	RA	UT	PB	SM	LP	SB
Empirical findings and diagnostic evaluation											
Lyu et al. (2025)	Consistency-based confidence estimation	✗	✗	✓	✗	✗	✗	✓	✗	✗	✓
Xiong et al. (2024)	Confidence elicitation and failure prediction	✗	✗	✓	✗	✗	✗	✓	✗	✗	✓
Tian et al. (2023)	Prompt-based elicitation from aligned models	✗	✗	✓	✗	✗	✗	✓	✗	✗	✗
Ni et al. (2024)	Probabilistic versus verbalized confidence	✗	✗	✓	✓	✗	✗	✓	✗	✓	✗
Yona et al. (2024)	Faithfulness of verbalized to intrinsic uncertainty	✗	✗	✓	✓	✗	✗	✗	✗	✗	✓
Zhang et al. (2024c)	Verbosity as a behavioral uncertainty signal	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗
Lin et al. (2024)	Graph-based black-box uncertainty quantification	✗	✗	✗	✗	✗	✗	✗	✓	✗	✓
Nikitin et al. (2024)	Semantic uncertainty from kernel similarities	✗	✗	✗	✗	✗	✗	✗	✓	✗	✓

Continued on next page

Table 1 (continued)

Work	Primary focus	FT	DC	FA	BA	RA	UT	PB	SM	LP	SB
Hou et al. (2024)	Input clarification ensembling for decomposition	✗	✗	✗	✗	✗	✗	✓	✗	✓	✓
Tomov et al. (2026)	Effect of ambiguity on UQ proxies	✗	✓	✓	✗	✗	✗	✗	✗	✓	✓
Yang et al. (2026)	Higher-order uncertainty via imprecise probabilities	✗	✗	✓	✗	✗	✗	✓	✗	✓	✗
Jiang et al. (2025)	Conformal calibration of factuality versus specificity	✗	✓	✓	✗	✗	✗	✗	✗	✓	✗
Zhang et al. (2024a)	Atomic (claim-level) calibration in long-form text	✗	✗	✓	✗	✗	✗	✓	✓	✓	✓
Zhou et al. (2024)	Linguistic uncertainty and its effect on human reliance	✗	✗	✓	✓	✗	✗	✓	✗	✗	✗
Training and alignment for uncertainty elicitation											
Lin et al. (2022a)	Supervised learning of calibrated verbal confidence	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗
Band et al. (2024)	Long-form linguistic calibration via distillation and RL	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗
Kapoor et al. (2024)	White-box correctness prediction with probes and LoRA	✓	✓	✗	✓	✓	✗	✓	✗	✗	✗
Zhang et al. (2024b)	Refusal-aware instruction tuning	✓	✓	✓	✗	✓	✗	✗	✗	✗	✗
Xu et al. (2024)	Confidence with self-reflective rationales	✓	✓	✓	✗	✗	✗	✗	✗	✗	✓
Liu et al. (2024)	Self-training from sampled and probabilistic uncertainty	✓	✓	✓	✗	✗	✗	✗	✓	✓	✓
Hager et al. (2025)	Distillation of semantic uncertainty into one pass	✓	✓	✓	✗	✗	✗	✗	✓	✗	✓
Tjandra et al. (2024)	Semantic-entropy supervision for learned abstention	✓	✓	✓	✗	✓	✗	✗	✓	✓	✓
Stengel-Eskin et al. (2024)	Listener-aware preference optimization of tone	✓	✓	✓	✓	✓	✗	✗	✗	✗	✗
Jang et al. (2025)	Scalar confidence supervision, emergent self-verification	✓	✓	✓	✓	✗	✗	✗	✗	✗	✓
Cohen et al. (2024)	Explicit uncertainty through an added [IDK] token	✓	✗	✓	✗	✓	✓	✗	✗	✓	✗
Li et al. (2025)	Proper-scoring-rule training for confidence tokens	✓	✗	✓	✗	✗	✓	✗	✗	✓	✗
Chuang et al. (2025)	Confidence tokens for routing and selective prediction	✓	✓	✓	✗	✗	✓	✗	✗	✓	✗
Downstream applications of uncertainty signals											
Manakul et al. (2023)	Sampling-based zero-resource hallucination detection	✗	✗	✗	✗	✗	✗	✓	✓	✗	✓
Quevedo et al. (2024)	Hallucination detection from token-probability features	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗
Farquhar et al. (2024)	Confabulation detection using semantic entropy	✗	✗	✗	✗	✗	✗	✗	✓	✓	✓
Chen et al. (2024)	Hallucination detection from internal states	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
Gupta et al. (2024)	Token-level uncertainty for cascade deferral	✗	✓	✗	✗	✓	✗	✗	✗	✓	✗
Ren et al. (2025)	Metacognitive control of test-time reasoning	✗	✗	✗	✓	✗	✗	✓	✗	✗	✗
Nguyen et al. (2026)	Aspect-based causal and interpretable abstention	✗	✗	✗	✓	✓	✗	✓	✗	✗	✗

Column definitions. **FT** (fine-tuning): modifies the model through supervised fine-tuning, reinforcement learning, preference optimization, LoRA, or continued pretraining. **DC** (dataset curation): constructs, partitions, filters, or automatically annotates a dedicated training, calibration, or preference dataset. **FA** (faithfulness): explicitly evaluates or optimizes whether expressed confidence, linguistic hedging, uncertainty tokens, or abstention behavior agrees with an underlying correctness or uncertainty signal. **BA** (behavioral analysis): analyzes linguistic behavior, indirect behavioral signals, metacognitive behavior, or human responses to uncertainty expressions. **RA** (refusal/abstention): introduces, trains, or evaluates refusal or abstention behavior rather than only a score that could later be thresholded. **UT** (uncertainty tokens): introduces a dedicated uncertainty or confidence token or vocabulary element rather than ordinary natural-language hedging. **PB** (prompt-based): uses prompting, elicitation, or multi-agent instructions as a central methodological component. **SM** (semantic metrics): uses meaning-level equivalence, entailment, or similarity to measure uncertainty. **LP** (logit/probabilistic): derives uncertainty from logits, token probabilities, or an explicit predictive probability representation. **SB** (sampling-based): estimates uncertainty from multiple generations, reasoning paths, clarifications, or model predictions.

Table 2: Practical requirements of the main method families. k denotes the number of sampled generations.

Family	Model access	Inference cost	Training cost	Representative works
Token-probability statistics	Logits	1 pass	None	Quevedo et al. (2024); Gupta et al. (2024)
Internal-state estimators	Hidden states	1 pass	None or a light probe	Chen et al. (2024)
Behavioural proxies	Black-box	1 pass	None	Zhang et al. (2024c)
Sampling / self-consistency	Black-box	k passes	None	Lyu et al. (2025); Xiong et al. (2024); Manakul et al. (2023)
Semantic entropy and kernel variants	Black-box + NLI model	k passes $+ O(k^2)$	None	Kuhn et al. (2023); Farquhar et al. (2024); Nikitin et al. (2024)
Similarity-graph estimators	Black-box	k passes $+ O(k^2)$	None	Lin et al. (2024)
Input clarification ensembling	Black-box	k clarified passes	None	Hou et al. (2024)
Verbalized prompting	Black-box	1–k passes	None	Tian et al. (2023); Yang et al. (2026)
Supervised or RL verbalization	White-box or API FT	1 pass	SFT/RL + curation	Lin et al. (2022a); Xu et al. (2024); Band et al. (2024); Liu et al. (2024); Jang et al. (2025)
Preference-based tone alignment	White-box	1 pass	DPO + listener simulation	Stengel-Eskin et al. (2024)
Dedicated uncertainty tokens	White-box (vocabulary)	1 pass	Continued pretraining or PEFT	Cohen et al. (2024); Li et al. (2025); Chuang et al. (2025)
Refusal-aware tuning	White-box	1 pass	SFT + knowledge partitioning	Zhang et al. (2024b); Tjandra et al. (2024)
Uncertainty distillation	White-box or API FT	1 pass	Offline sampling + calibration	Hager et al. (2025)
Post-hoc conformal calibration	Black-box + held-out set	1 pass + calibration	Calibration data only	Jiang et al. (2025)
Metacognitive test-time control	Black-box	Variable, adaptive	None	Ren et al. (2025)
Agentic aspect-based probing	Black-box	Many passes + agents	None	Nguyen et al. (2026)

7 Open Problems and Future Directions

Despite the progress reviewed above, several limitations prevent current uncertainty-aware methods from providing a reliable and general-purpose capability. These limitations are closely connected: current estimators often conflate different sources of uncertainty, expensive signals are difficult to deploy, and training methods may learn behaviors that do not transfer beyond the conditions under which they were trained.

Separating ambiguity from lack of knowledge. A fundamental limitation is that most existing methods reduce uncertainty to a single scalar, even though uncertainty can arise for different reasons. A model may be uncertain because it lacks the knowledge required to answer a question, or because the question itself admits multiple valid interpretations or answers. Sampling- and consistency-based estimators can treat both cases as disagreement, although they require different system responses: missing knowledge may call for retrieval or abstention, whereas ambiguity may call for clarification or presentation of multiple valid alternatives (Hou et al., 2024; Tomov et al., 2026). Future benchmarks and training objectives should therefore evaluate these sources separately and provide supervision that distinguishes input ambiguity from uncertainty about the answer.

From expensive estimation to efficient uncertainty awareness. A second limitation concerns the computational cost of the strongest uncertainty estimators. Semantic and graph-based methods often require multiple generations followed by semantic comparison, while clarification-based approaches require additional

model calls. These methods can provide substantially richer signals, but their cost becomes problematic when uncertainty is itself intended to reduce inference cost through routing, filtering, or adaptive computation. Existing distillation approaches address this problem by moving the expensive computation offline (Hager et al., 2025), but this creates a dependency on the quality and distribution of the offline estimator. A key direction is therefore to develop uncertainty representations that retain the benefits of semantic or sampling-based estimation while producing reliable signals in a single forward pass.

Robustness beyond the training distribution. Current results are also concentrated on relatively narrow benchmark settings, and it remains unclear how reliably learned uncertainty transfers across domains, subjects, formats, and model families. A confidence signal may be well calibrated on the distribution used to construct its training data while becoming overconfident or excessively conservative under distribution shift. This issue is particularly important for methods that learn a specific uncertainty vocabulary or abstention policy, because the learned behavior may reflect properties of the training distribution rather than a general representation of the model's knowledge boundary. Future evaluations should therefore treat out-of-distribution calibration and threshold transfer as primary evaluation targets rather than secondary analyses.

Beyond binary abstention. Most uncertainty-aware systems ultimately reduce the decision to answering or refusing. This binary formulation discards a useful middle ground in which the model can reduce the specificity of its claim while remaining informative. Conformal linguistic calibration provides an initial step in this direction by allowing response specificity to vary with confidence (Jiang et al., 2025). A more general framework should allow the system to choose among precise answers, qualified answers, clarification, retrieval, additional reasoning, escalation, and abstention according to both uncertainty and application-specific error costs. Such a framework would turn uncertainty from a reporting variable into a decision-making signal.

Uncertainty in multi-step and tool-augmented systems. Finally, most existing evaluations consider isolated question-answering settings, whereas practical LLM systems increasingly retrieve information, invoke tools, interact over multiple turns, and route requests between models. In these settings, uncertainty is no longer associated with a single generation. The system must reason about uncertainty in its parametric knowledge, retrieved evidence, intermediate reasoning steps, and external tools, and must decide when additional computation is justified. An important open direction is therefore to develop uncertainty-aware controllers that map these signals to actions such as retrieval, additional reasoning, model escalation, or abstention while explicitly accounting for the cost of each action.

Understanding the underlying mechanism. These limitations ultimately point to a deeper question: whether current methods teach models to understand their own uncertainty or merely teach them to imitate calibrated uncertainty behavior. Existing work provides evidence that uncertainty-related representations can be associated with separable internal states and can improve after uncertainty supervision (Lin et al., 2022a; Kapoor et al., 2024), but the mechanism remains poorly understood. Progress in this direction requires studying how uncertainty information is represented during pretraining, how it changes during alignment, and whether the resulting representation corresponds to a genuine estimate of knowledge reliability rather than a learned linguistic policy.

Taken together, these gaps suggest a shift in the research objective. Rather than optimizing uncertainty estimation as an isolated prediction problem, future work should treat uncertainty awareness as a capability that must remain meaningful across different sources of uncertainty, computational budgets, distributions, and downstream decisions. Such a view would connect uncertainty estimation with the broader goal of building systems that know when to answer, when to qualify an answer, and when to take a different action.

8 Conclusion

This survey examined uncertainty self-awareness in large language models from three complementary perspectives: whether models can internally estimate the limits of their knowledge, whether they can communicate these estimates faithfully, and whether such signals can be used to improve inference-time decisions.

Across the literature, a consistent picture emerges: although useful uncertainty information is often present in models, it is imperfectly communicated, with systematic overconfidence further amplified by instruction tuning and preference optimization. More sophisticated estimators, particularly those based on repeated generations or semantic agreement, can provide stronger uncertainty signals but introduce substantial computational costs, creating a fundamental trade-off between estimation quality and practical deployment. The field is further hindered by inconsistent definitions of uncertainty, the difficulty of distinguishing ambiguity from ignorance, response-level evaluation that obscures claim-level miscalibration, limited robustness to distribution shifts, and benchmarks that are predominantly English, short-form, single-answer, and increasingly contaminated by model pretraining. Existing approaches can therefore be viewed as addressing three closely related but distinct challenges—uncertainty estimation, uncertainty communication, and uncertainty-aware decision making—whose improvements do not necessarily transfer across one another. Future research should consequently move beyond post-hoc calibration toward treating uncertainty awareness as a capability learned throughout training, develop richer and ambiguity-aware uncertainty representations, evaluate models under controlled inference budgets and at the claim level, and use uncertainty as an active control signal for retrieval, abstention, adaptive reasoning, and model routing. Ultimately, the goal is not simply to make models less confident when they are wrong, but to enable them to reliably distinguish between what they know, what they do not know, and what is inherently uncertain, and to act and communicate accordingly.

References

Michael Ahn et al. Do as i can, not as i say: Grounding language in robotic affordances. In *Conference on Robot Learning*, 2022.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, et al. LongBench: A bilingual, multitask benchmark for long context understanding. *arXiv preprint arXiv:2308.14508*, 2023.

Neil Band, Xuechen Li, Tengyu Ma, and Tatsunori Hashimoto. Linguistic calibration of long-form generations. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024. arXiv:2404.00474.

Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1): 1–3, 1950.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. INSIDE: LLMs' internal states retain the power of hallucination detection. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. arXiv:2402.03744.

Yu-Neng Chuang, Prathusha K Sarma, John Boccio, Sara Bolouki, Xia Hu, Parikshit Gopalan, and Helen Zhou. Learning to route llms with confidence tokens. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Roi Cohen, Konstantin Dobler, Eden Biran, and Gerard de Melo. I don't know: Explicit modeling of uncertainty with an [idk] token. In *Advances in Neural Information Processing Systems*, 2024. URL `https://arxiv.org/abs/2412.06676`.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. A dataset of information-seeking questions and answers anchored in research papers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics*, 2021.

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

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. volume 630, pp. 625–630. Nature Publishing Group UK London, 2024.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, 2017.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. A survey of confidence estimation and calibration in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6577–6595, 2024.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361, 2021.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, 2017.

Neha Gupta, Harikrishna Narasimhan, Wittawat Jitkrittum, Ankit Singh Rawat, Aditya Krishna Menon, and Sanjiv Kumar. Language model cascades: Token-level uncertainty and beyond. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. arXiv:2404.10136.

Sophia Hager, David Mueller, Kevin Duh, and Nicholas Andrews. Uncertainty distillation: Teaching language models to express semantic confidence. *arXiv preprint arXiv:2503.14749*, 2025.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, 2020.

Bairu Hou, Yujian Liu, Kaizhi Qian, Jacob Andreas, Shiyu Chang, and Yang Zhang. Decomposing uncertainty for large language models through input clarification ensembling, 2024. URL `https://arxiv.org/abs/2311.08718`.

Shengding Hu, Yifan Luo, Huadong Wang, Xingyi Cheng, Zhiyuan Liu, and Maosong Sun. Won't get fooled again: Answering questions with false premises. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 2023.

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

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. *ACM Transactions on Information Systems*, 43(2): 1–55, 2025.

Chaeyun Jang, Moonseok Choi, Seungyoo Lee, Yegon Kim, Hyungi Lee, and Juho Lee. Verbalized confidence triggers self-verification: Emergent behavior without explicit reasoning supervision. *arXiv preprint arXiv:2506.03723*, 2025.

Ziwei Ji, Delong Chen, Etsuko Ishii, Samuel Cahyawijaya, Yejin Bang, Bryan Wilie, and Pascale Fung. Llm internal states reveal hallucination risk faced with a query. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 88–104, 2024.

Zhengping Jiang, Anqi Liu, and Benjamin Van Durme. Conformal linguistic calibration: Trading-off between factuality and specificity, 2025. URL `https://arxiv.org/abs/2502.19110`.

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

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pp. 1601–1611, 2017.

Ryo Kamoi, Tanya Goyal, Juan Diego Rodriguez, and Greg Durrett. WiCE: Real-world entailment for claims in wikipedia. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

Sanyam Kapoor, Nate Gruver, Manley Roberts, Katherine Collins, Arka Pal, Umang Bhatt, Adrian Weller, Samuel Dooley, Micah Goldblum, and Andrew Gordon Wilson. Large language models must be taught to know what they don't know. In *Thirty-Eighth Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

Sunnie SY Kim, Jennifer Wortman Vaughan, Q Vera Liao, Tania Lombrozo, and Olga Russakovsky. Fostering appropriate reliance on large language models: The role of explanations, sources, and inconsistencies. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pp. 1–19, 2025.

Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. The NarrativeQA reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328, 2018.

Anastasia Krithara, Anastasios Nentidis, Konstantinos Bougiatiotis, and Georgios Paliouras. BioASQ-QA: A manually curated corpus for biomedical question answering. *Scientific Data*, 10(1):170, 2023.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023. arXiv:2302.09664.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, et al. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019.

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. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

Yibo Li, Miao Xiong, Jiaying Wu, and Bryan Hooi. Conftuner: Training large language models to express their confidence verbally. In *Advances in Neural Information Processing Systems*, 2025. URL `https://arxiv.org/abs/2508.18847`.

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words. *arXiv preprint arXiv:2205.14334*, 2022a.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 2022b.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Generating with confidence: Uncertainty quantification for black-box large language models, 2024. URL `https://arxiv.org/abs/2305.19187`.

Shudong Liu, Zhaocong Li, Xuebo Liu, Runzhe Zhan, Derek F Wong, Lidia S Chao, and Min Zhang. 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*, pp. 21635–21645, 2024.

Xiaoou Liu, Tiejin Chen, Longchao Da, Chacha Chen, Zhen Lin, and Hua Wei. 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 (KDD)*, pp. 6107–6117, 2025. doi: 10.1145/3711896.3736569.

Qing Lyu, Kumar Shridhar, Chaitanya Malaviya, Li Zhang, Yanai Elazar, Niket Tandon, Marianna Apidianaki, Mrinmaya Sachan, and Chris Callison-Burch. Calibrating large language models with sample consistency. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 19260–19268, 2025. doi: 10.1609/aaai.v39i18.34120.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 2023.

Potsawee Manakul, Adian Liusie, and Mark Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9004–9017, 2023.

Shen-Yun Miao, Chao-Chun Liang, and Keh-Yih Su. A diverse corpus for evaluating and developing english math word problem solvers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. AmbigQA: Answering ambiguous open-domain questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 2020.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 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*, 2023.

Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2015.

Vy Nguyen, Ziqi Xu, Jeffrey Chan, Estrid He, Feng Xia, and Xiuzhen Zhang. Hallucinate less by thinking more: Aspect-based causal abstention for large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 32555–32563, 2026.

Shiyu Ni, Keping Bi, Lulu Yu, and Jiafeng Guo. Are large language models more honest in their probabilistic or verbalized confidence? In *China Conference on Information Retrieval*, pp. 124–135. Springer, 2024.

Alexander Nikitin, Jannik Kossen, Yarin Gal, and Pekka Marttinen. Kernel language entropy: Fine-grained uncertainty quantification for llms from semantic similarities, 2024. URL `https://arxiv.org/abs/2405.20003`.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. Are NLP models really able to solve simple math word problems? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics*, 2021.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019.

Ernesto Quevedo, Pablo Rivas, Jorge Yero Salazar, Rachel Koerner, and Tomas Cerny. Detecting hallucinations in large language model generation: A token probability approach. In *World Congress in Computer Science, Computer Engineering & Applied Computing*, pp. 154–173. Springer, 2024.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392, 2016.

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

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

Allen Ren et al. Llms know when they know, but do not act on it: A metacognitive harness for test-time scaling. *arXiv preprint*, 2025.

Subhro Roy and Dan Roth. Solving general arithmetic word problems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. WinoGrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106, 2021.

Koustuv Sinha, Shagun Sodhani, Jin Dong, Joelle Pineau, and William L. Hamilton. CLUTRR: A diagnostic benchmark for inductive reasoning from text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019.

Aarohi Srivastava et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, 2023.

Elias Stengel-Eskin, Peter Hase, and Mohit Bansal. LACIE: Listener-aware finetuning for confidence calibration in large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=RnvgYd9RAh`.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, 2019.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: A large-scale dataset for fact extraction and verification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics*, 2018.

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. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5433–5442, 2023.

Benedict Aaron Tjandra, Muhammed Razzak, Jannik Kossen, Kunal Handa, and Yarin Gal. Fine-tuning large language models to appropriately abstain with semantic entropy, 2024. URL `https://arxiv.org/abs/2410.17234`.

Tim Tomov, Dominik Fuchsgruber, Tom Wollschläger, and Stephan Günnemann. The illusion of certainty: Uncertainty quantification for llms fails under ambiguity. *arXiv preprint arXiv:2511.04418*, 2026.

SM Tonmoy, SM Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv preprint arXiv:2401.01313*, 2024.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022.

Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, et al. Long-form factuality in large language models. In *Advances in Neural Information Processing Systems*, 2024.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, 2017.

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*, 2024.

Tianyang Xu, Shujin Wu, Shizhe Diao, Xiaoze Liu, Xingyao Wang, Yangyi Chen, and Jing Gao. Sayself: Teaching llms to express confidence with self-reflective rationales, 2024. URL `https://arxiv.org/abs/2405.20974`.

Anita Yang, Krikamol Muandet, Michele Caprio, Siu Lun Chau, and Masaki Adachi. Verbalizing llm's higher-order uncertainty via imprecise probabilities. *arXiv preprint arXiv:2603.10396*, 2026.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380, 2018.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. Do large language models know what they don't know? In *Findings of the Association for Computational Linguistics: ACL 2023*, 2023.

Gal Yona, Roee Aharoni, and Mor Geva. Can large language models faithfully express their intrinsic uncertainty in words? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 7752–7764, 2024.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

Caiqi Zhang, Ruihan Yang, Zhisong Zhang, Xinting Huang, Sen Yang, Dong Yu, and Nigel Collier. Atomic calibration of llms in long-form generations. *arXiv preprint arXiv:2410.13246*, 2024a.

Rui Zhang et al. R-tuning: Instructing large language models to say "i don't know". *arXiv preprint*, 2024b.

Yusen Zhang, Sarkar Snigdha Sarathi Das, and Rui Zhang. Verbosity $\neq$ veracity: Demystify verbosity compensation behavior of large language models, 2024c. URL `https://arxiv.org/abs/2411.07858`.

Wenting Zhao, Tanya Goyal, Yu Ying Chiu, Liwei Jiang, et al. WildHallucinations: Evaluating long-form factuality in LLMs with real-world entity queries. In *arXiv preprint arXiv:2407.17468*, 2024.

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

A Background

As briefly introduced in Section 2 of the main text, defining and quantifying uncertainty in large language models requires a precise theoretical framework. This appendix provides a comprehensive and formal treatment of the vocabulary, mathematical formulations, and evaluation metrics used throughout this survey.

In the following subsections, we expand upon the theoretical distinction between aleatoric and epistemic uncertainty, provide the rigorous mathematical formulations for confidence elicitation (including graph Laplacian constructions for semantic uncertainty), and present the exact equations for the metrics used to evaluate these estimators.

A.1 Aleatoric and Epistemic Uncertainty

Uncertainty in machine learning is traditionally split into two kinds, and the distinction matters here more than usual, because the two call for different actions.

- **Aleatoric uncertainty** arises from ambiguity, noise, or genuine multiplicity in the data itself. It is irreducible: even a perfect model trained on infinite data cannot remove it. In language, this covers questions that legitimately admit several correct answers, prompts that are underspecified ("Who is the president of this country?"), and open-ended generation where many outputs are equally valid.
- **Epistemic uncertainty** arises from the limits of the model's own knowledge. It reflects what the model *does not know*, and it can in principle be reduced with more data, better training, or retrieval. A question about an event after the model's knowledge cutoff produces purely epistemic uncertainty.

In the context of LLMs, most work on self-awareness targets epistemic uncertainty, estimating what the model itself does not know so that it can abstain or hedge instead of fabricating an answer. Recent work shows that aleatoric uncertainty matters too, since ignorance and input ambiguity produce the same disagreement among sampled answers yet call for different actions: an ignorant model should retrieve or abstain, whereas an ambiguous question should be sent back to the user for clarification, or answered with a broader claim the model can actually support (Tomov et al., 2026; Hou et al., 2024; Jiang et al., 2025). Estimating both types is therefore a prerequisite for choosing the right action, not a side concern.

A.2 Confidence, Uncertainty, and Calibration

There is no unified definition of "uncertainty" and "confidence" in the LLM uncertainty-estimation literature, which is a genuine source of confusion rather than a stylistic matter: two papers reporting a quantity under the same name are not always estimating the same thing. Below we describe the two most widely used conventions.

A.2.1 Uncertainty as the Complement of Confidence

The most common convention treats the two as complementary views of a single quantity. **Confidence** is the model's quantified belief that its generated output is correct, derived either from the output token distribution or produced explicitly as text, and **uncertainty** is simply its complement,

$$U = 1 - C, \tag{1}$$

where C denotes the model's confidence. Under this reading a single number is expected to answer two questions at once: how hard or ill-posed the question is, and how likely this particular answer is to be correct.

A.2.2 Uncertainty as a Property of the Input, Confidence as a Property of the Output

The second convention separates the two deliberately. Confidence is an output-dependent quantity describing the model's belief in one particular generated response, whereas uncertainty is an input-dependent property reflecting the ambiguity or difficulty of the prompt itself, which therefore cannot be inferred from the confidence assigned to a single output.

Lin et al. (2024) give this separation a formal statement and show how both quantities can be obtained from one construction, without access to token probabilities. Given m responses sampled for the same prompt, they build a weighted graph whose nodes are the responses and whose edge weights $w_{ij} \in [0, 1]$ record the semantic similarity of s_i and s_j, typically taken from the entailment probabilities of a natural language inference model. With the degree matrix $D_{ii} = \sum_j w_{ij}$, the symmetric normalised graph Laplacian is

$$L = I - D^{-1/2} W D^{-1/2}. \tag{2}$$

The spectrum of this graph gives the input-level quantity. For a graph with binary edges, the number of connected components which equivalently is the multiplicity of the eigenvalue zero of L, is exactly the number of distinct meanings the model produced for that prompt. Continuous similarities make a graph that is almost always fully connected, leaving a single zero eigenvalue, so the count is relaxed to the small eigenvalues,

$$U_{\mathrm{EigV}}(x) = \sum_{k=1}^{m} \max(0, 1 - \lambda_k), \tag{3}$$

which acts as a continuous estimate of the number of semantic meanings and serves as the uncertainty of the query. The *nodes* give the output-level quantity: since each node is one response, its degree D_{jj} measures how strongly that response is supported by the others, so a well-connected node lies in a confident region of the output distribution and a peripheral one does not. One graph thus yields a single uncertainty value for the question and a separate confidence value for every candidate answer.

A.2.3 Calibration

Whichever convention is adopted, a confidence score is only useful if the number the model reports matches how likely it actually is to be right. Aligning the two is called **calibration**. Formally, a model is perfectly calibrated if, for every confidence level $p \in [0, 1]$, the probability that its prediction is correct is exactly p:

$$\mathbb{P}(\hat{Y} = Y \mid \hat{P} = p) = p, \tag{4}$$

where Y is the ground-truth answer, $\hat{Y}$ the model's prediction, and $\hat{P}$ its stated confidence. In other words, among all the answers a calibrated model reports with 70% confidence, about seventy percent should be correct. The gap between the two is the calibration error, quantified with metrics such as Expected Calibration Error (Section A.4), and it takes two forms:

- **Overconfidence:** stated confidence systematically exceeds the true accuracy rate. This is by far the dominant failure mode in LLMs, and it is often exacerbated by RLHF alignment, which penalises hedged language in preference data (Zhou et al., 2024).
- **Underconfidence:** stated confidence falls below the model's true empirical accuracy.

A.3 How Uncertainty Is Elicited from a Model

Uncertainty must first be extracted before it can be evaluated or used. Existing methods fall into three modalities that differ in what access to the model they assume.

A.3.1 Logit-Based (Intrinsic Probabilistic) Uncertainty

Logit-based methods read uncertainty directly from the predictive distribution produced during decoding, and therefore require open weights or an API that exposes token probabilities. Given a prompt X and a generated sequence $Y = (y_1, \ldots, y_t)$, the model defines $P(y_t \mid y_{<t}, X)$ over the vocabulary at each step. The standard token-level measure is the Shannon entropy of that distribution,

$$H(y_t \mid y_{<t}, X) = -\sum_{w \in \mathcal{V}} P(y_t = w \mid y_{<t}, X) \log P(y_t = w \mid y_{<t}, X), \tag{5}$$

where $\mathcal{V}$ is the vocabulary and higher entropy means a flatter, less committed distribution. Other common proxies are the maximum token probability, the sequence log-likelihood, the average token probability, and the minimum token probability over the generated sequence,

$$mtp = \min_i P(y_i). \tag{6}$$

These estimators cost only a single forward pass, which is their main attraction, but they operate on tokens rather than meanings and are unavailable for closed commercial models.

A.3.2 Sampling-Based Uncertainty

Sampling-based methods generate several responses to the same prompt under stochastic decoding and infer confidence from how much those responses agree. The intuition is that a model that keeps returning the same answer is confident, while one that returns contradictory answers is not. Representative variants are self-consistency, which samples independent reasoning paths and measures agreement among the final answers (Lyu et al., 2025); semantic entropy, which first groups responses into semantically equivalent clusters using an NLI model and then computes entropy over the clusters, making the estimate invariant to paraphrase (Kuhn et al., 2023; Farquhar et al., 2024); and sample dispersion measures based on embedding variance or on the similarity graph of Section A.2 (Lin et al., 2024). These methods capture variability that a single forward pass cannot, at the cost of k generations and, for the semantic variants, $O(k^2)$ comparisons.

A.3.3 Verbalized (Linguistic) Confidence

Verbalized methods move the elicitation out of probability space and into language, which makes them applicable to black-box systems. They take three forms. In numerical scoring, the model is prompted to attach a number to its answer ("Answer the question and provide your confidence level from 0% to 100%"). In linguistic hedging, uncertainty is conveyed through wording such as "I am not entirely sure, but…" or an explicit "I don't know." In post-hoc self-probing, the model is asked in a second pass whether its own answer $\hat{Y}$ to X is correct, and the probability of the token "Yes", or the proportion of affirmative responses across trials, is used as the confidence score.

Verbalized approaches need no access to model internals and are the only option for proprietary APIs, but empirical studies consistently find them poorly calibrated and highly sensitive to prompt wording (Xiong et al., 2024; Ni et al., 2024).

Overall, the three modalities trade off differently: logit-based estimates are cheap but require probabilities and ignore meaning; sampling-based estimates are the most informative but the most expensive; verbalized estimates are universally deployable but the least reliable.

A.4 Metrics for Evaluating Uncertainty Estimates

Evaluating an uncertainty estimator is not the same as evaluating accuracy: the question is whether the estimate tracks the probability of being correct. Each metric below was introduced to capture something the previous one could not, and reading them in that order explains why the literature reports so many at once.

A.4.1 Expected Calibration Error

The oldest question is simply whether stated confidence matches empirical accuracy. Expected Calibration Error (ECE), introduced by Naeini et al. (2015) and popularised for neural networks by Guo et al. (2017), answers it by partitioning predictions into M confidence bins and averaging the gap between accuracy and confidence within each bin,

$$ECE = \sum_{m=1}^{M} \frac{|B_m|}{N} \left|\mathrm{acc}(B_m) - \mathrm{conf}(B_m)\right|, \tag{7}$$

where B_m is the m-th bin, $\mathrm{acc}(B_m)$ its empirical accuracy and $\mathrm{conf}(B_m)$ its average confidence. Lower is better. Two refinements target its weak points: **Maximum Calibration Error** (MCE) reports the largest single-bin gap rather than the average, which matters when a rare but badly miscalibrated region is the operational risk, and the Brier score (Brier, 1950) avoids binning altogether by scoring each prediction individually,

$$BS = \frac{1}{N} \sum_{i=1}^{N} (\hat{P}_i - y_i)^2, \qquad y_i \in \{0, 1\}, \tag{8}$$

with **negative log-likelihood** (NLL) serving the same purpose under a logarithmic penalty.

A.4.2 Discrimination: AUROC and AURC

ECE has a defect that matters in practice: a model can be well calibrated and still useless. Xiong et al. (2024) give the concrete example of GPT-4 on GSM8K, which reaches a near-optimal ECE simply by assigning 100% confidence to every sample, since its accuracy is around 93%; because all samples receive the same score, correct and incorrect answers cannot be told apart. What is missing is *discrimination*, and it is measured by the Area Under the ROC Curve (AUROC), which treats error prediction as binary classification and reports the probability that a randomly chosen incorrect prediction is assigned higher uncertainty than a randomly chosen correct one,

$$AUROC = P(U_{\text{incorrect}} > U_{\text{correct}}). \tag{9}$$

AUROC is the metric of choice for hallucination detection and confidence-based filtering. It is, however, threshold-free, and a deployed system must pick a threshold. Area Under the Risk–Coverage Curve (AURC) (Geifman & El-Yaniv, 2017) closes that gap by sorting predictions by confidence and tracking the error rate as the least confident samples are removed; lower AURC means the estimator successfully identifies unreliable predictions across operating points.

A.4.3 Selective Prediction and Abstention

Once a model is allowed to decline, the trade-off is between how much it answers and how often it is right. Given a threshold, the model answers only above it and rejects the rest, and the resulting behaviour is summarised by the risk–coverage curve

$$Risk(c) = \mathbb{P}(\hat{Y} \neq Y \mid \text{coverage} = c), \tag{10}$$

where c is the fraction of retained predictions. A good estimator keeps risk low while keeping coverage high.

Risk–coverage curves describe a whole family of operating points, which is exactly what makes them awkward when one wants to compare two abstaining models with a single number. Tjandra et al. (2024) therefore proposed the Accuracy–Engagement Distance (AED), which measures how far a model sits from the ideal corner of perfect accuracy and maximum willingness to answer,

$$AED = \sqrt{\frac{I^2 + (|D| - C)^2}{2|D|^2}}, \tag{11}$$

where I is the number of incorrect responses, C the number of answered queries, and $|D|$ the total number of queries. Lower AED rewards a model that answers accurately while abstaining on the cases it cannot handle, and penalises both a model that answers everything wrongly and one that refuses everything. Its limitation is the mirror image of its convenience: it fixes a single exchange rate between a wrong answer and a refusal, whereas the correct rate is a property of the application, a point we develop in Section 7.

A.4.4 From Token-Level to Meaning-Level Measures

The classical estimators reduce uncertainty to token-level statistics, multiplying or averaging the probabilities of individual tokens or taking the entropy of the vocabulary distribution at each decoding step. Their common weakness is that they look at a response piece by piece and never at what it actually says: the same answer phrased in two ways yields two very different token sequences, so lexical variation is counted as disagreement while two genuinely contradictory answers of similar length may score alike.

Graph-based methods address this by treating the response as a whole. For a given question, several answers are sampled and arranged as the nodes of a graph whose edges record how semantically close any two of them are, so that the spread of the answers rather than the shape of any single token distribution becomes the object being measured (Lin et al., 2024; Nikitin et al., 2024). As shown in Section A.2, this construction is also what made the second definition of uncertainty and confidence possible, since the graph separates a question-level quantity, read from its overall connectivity, from a response-level quantity, read from each individual node.

Measuring dispersion on such a graph requires a different notion of entropy, because continuous edge weights leave no discrete distribution over clusters for Shannon entropy to apply to. The graph is therefore turned into a normalised kernel that behaves like a density matrix, and its entropy is taken in the Von Neumann sense, the matrix generalisation of Shannon entropy developed in quantum theory. Such measures reduce to ordinary semantic entropy (Kuhn et al., 2023; Farquhar et al., 2024) when responses do fall into cleanly separated meaning clusters, and give a finer-grained account of semantic spread when they do not.

A.4.5 Higher-Order Uncertainty

All of the discussed methods return a single number as a number, which presupposes that the model has one well-determined distribution over answers and only needs to summarise it. A recent line of work argues that this presupposition fails precisely where uncertainty matters most, and separates two orders (Yang et al., 2026; Tomov et al., 2026). First-order uncertainty is uncertainty over the possible answers, which is what all the estimators above report. Second-order uncertainty is uncertainty about that distribution itself in the case where the model cannot even settle on how likely its own answers are. The two correspond loosely, though not exactly, to the aleatoric and epistemic categories of Section **??**.

Capturing the second order requires a representation wider than a point estimate, so confidence is expressed as a probability interval rather than a single value, with the width of the interval measuring how indeterminate the belief is. This separates two situations that a scalar collapses into one: believing an answer is probably wrong, and having no basis for assigning a probability to it at all.

B Datasets and Benchmarks

As introduced in the main text, the empirical claims in the uncertainty-awareness literature heavily rely on benchmarks originally designed for standard question answering, reading comprehension, or reasoning tasks. These datasets were later repurposed because they provide the ground-truth answers necessary for scoring confidence estimates.

This appendix provides a comprehensive and detailed breakdown of all the datasets referenced throughout the survey. Below, we describe each dataset in depth, grouped by the specific role it plays in uncertainty evaluation, and outline the unique characteristics and limitations that constrain what can be concluded from them. A complete summary of these benchmarks is also provided in Table 3 at the end of this section.

B.1 Short-Form Factual Question Answering

The core of the field is closed-book factual QA with short answers, because correctness can be checked automatically and a single answer string makes semantic clustering tractable.

TriviaQA (Joshi et al., 2017) contains over 650K question–answer–evidence triples collected from trivia websites. Used closed-book, with the evidence documents discarded, it becomes a clean test of parametric knowledge and is the single most common dataset in this literature, appearing in semantic entropy (Farquhar et al., 2024), KLE (Nikitin et al., 2024), black-box graph estimators (Lin et al., 2024), LACIE (Stengel-Eskin et al., 2024), IDK-tuning (Cohen et al., 2024), and ConfTuner (Li et al., 2025). Its advantage is scale and answer-string matching; its drawback is that the questions are mostly popular-entity trivia, so accuracy is high and the interesting low-confidence region is thin.

Natural Questions and its open-domain variant **NQ-Open** (Kwiatkowski et al., 2019) consist of real anonymised Google search queries paired with Wikipedia pages. Because the questions were issued by users rather than written by annotators, they are harder and messier than TriviaQA, which is precisely why Ni et al. (2024) chose NQ as their main benchmark for comparing probabilistic and verbalized confidence, and why Yona et al. (2024) use it to study faithfulness. The same realism is also a liability: a substantial fraction of NQ-Open questions are ambiguous, which means disagreement among samples is not always evidence of ignorance.

SQuAD (Rajpurkar et al., 2016) is an extractive reading-comprehension benchmark of crowd-sourced questions over Wikipedia paragraphs, with answers as spans of the provided context. In uncertainty work it is

used mainly in the open-book setting to produce longer, descriptive answers than TriviaQA does (Farquhar et al., 2024; Nikitin et al., 2024), and its F1 metric is frequently reused to score free-form generations.

PopQA (Mallen et al., 2023) contains entity-centric factual questions whose subject entities are associated with Wikipedia page-view statistics, providing a proxy for how common or rare the underlying facts are. This makes the dataset useful for studying uncertainty across different levels of factual familiarity where models can be evaluated not only on whether they answer correctly, but also on whether their confidence or uncertainty changes appropriately as questions move from common to long-tail knowledge. Related dataset constructions create similar frequency contrasts by comparing questions about more and less familiar entities (Yona et al., 2024; Cohen et al., 2024; Ni et al., 2024).

CoQA (Reddy et al., 2019) and **SciQ** (Welbl et al., 2017) provide additional variation beyond standard factual QA. CoQA contains conversational questions with free-form answers and dependencies across dialogue turns, making it useful for evaluating uncertainty in a conversational setting (Kuhn et al., 2023; Lin et al., 2024). SciQ consists of science exam questions and has been used both for training uncertainty-related predictors and for evaluating their transfer beyond the domains seen during training (Kapoor et al., 2024; Band et al., 2024).

B.2 Domain-Specific Question Answering

BioASQ (Krithara et al., 2023) is a biomedical question-answering benchmark built around expert information needs and includes yes/no, factoid, and longer summary-style answers. Its specialized domain makes it useful for testing whether uncertainty estimates developed on general-domain QA remain reliable when the model is evaluated on less familiar and more technical knowledge (Farquhar et al., 2024; Nikitin et al., 2024; Band et al., 2024). **MedQA** (Jin et al., 2021) consists of multiple-choice questions derived from medical licensing examinations. It provides a complementary setting in which models must reason over specialized medical knowledge within a fixed candidate set, allowing uncertainty methods to be evaluated under domain shift without the additional ambiguity of open-ended answer generation (Chuang et al., 2025).

B.3 Reasoning and Mathematical Problem Solving

Reasoning benchmarks matter here because uncertainty may arise not only from whether the model possesses the relevant facts, but also from whether it can reliably carry out the intermediate steps needed to reach an answer. They therefore provide a setting in which confidence in the final prediction depends on the stability of a multi-step derivation rather than on factual recall alone.

GSM8K (Cobbe et al., 2021) contains grade-school math word problems requiring multi-step arithmetic, and **SVAMP** (Patel et al., 2021) contains adversarially perturbed word problems designed to defeat shallow pattern matching. Both are used by Xiong et al. (2024), and SVAMP is the one reasoning task included in the semantic entropy evaluations (Farquhar et al., 2024; Nikitin et al., 2024). **ASDiv** (Miao et al., 2020) and **MultiArith** (Roy & Roth, 2015) extend the same family and are used by Lyu et al. (2025).

StrategyQA (Geva et al., 2021) requires implicit multi-hop reasoning to answer yes/no questions, and several **BIG-bench** tasks (Srivastava et al., 2023), in particular Date Understanding, Object Counting, and Sports Understanding, provide symbolic and commonsense reasoning with unambiguous answers. Xiong et al. (2024) and Lyu et al. (2025) use these to show that verbalized confidence behaves differently on reasoning than on knowledge tasks. Lyu et al. (2025) additionally cover planning with **SayCan** (Ahn et al., 2022) and relational inference with **CLUTRR** (Sinha et al., 2019), giving one of the broadest task ranges in the field.

HotpotQA (Yang et al., 2018) is a Wikipedia-based multi-hop QA dataset in which the answer requires combining several passages. It is the in-distribution training set for SaySelf (Xu et al., 2024) and Conf-Tuner (Li et al., 2025) and one of the tasks in R-Tuning (Zhang et al., 2024b), chosen because multi-step questions produce a wider spread of correctness than single-hop trivia.

B.4 Multiple-Choice Knowledge Benchmarks

MMLU (Hendrycks et al., 2021) spans 57 subjects in multiple-choice form and is used in three distinct ways: as a whole for refusal-aware tuning (Zhang et al., 2024b) and routing (Chuang et al., 2025); through individual subsets such as Professional Law and Business Ethics, chosen by Xiong et al. (2024) as domains where models are expected to be least calibrated; and as a knowledge-retention check to confirm that uncertainty training has not damaged the underlying model (Cohen et al., 2024). **ARC** (Clark et al., 2018), **OpenBookQA** (Mihaylov et al., 2018), **CommonsenseQA** (Talmor et al., 2019), **HellaSwag** (Zellers et al., 2019), and **WinoGrande** (Sakaguchi et al., 2021) appear as the training pool for general-purpose correctness predictors (Kapoor et al., 2024) and as retention benchmarks (Cohen et al., 2024).

The multiple-choice format is convenient because a token probability over four options is directly a confidence score, but it also inflates apparent calibration where guessing among four options gives a floor accuracy of 25%, and the model never has to decide that no answer is available.

B.5 Truthfulness, Hallucination, and Refusal Benchmarks

A second family of datasets was constructed specifically to expose failure modes that ordinary question-answering benchmarks often underrepresent. Rather than sampling questions from general knowledge distributions, these datasets deliberately target situations in which models are prone to hallucinate, follow false premises, reproduce common misconceptions, or answer questions that should instead trigger refusal or uncertainty. They are therefore particularly useful for evaluating whether a model can recognize unreliable or unanswerable cases rather than merely achieve high average accuracy.

TruthfulQA (Lin et al., 2022b) collects questions on which humans commonly hold misconceptions, so that imitation of the training corpus produces confident falsehoods. It is used as the transfer benchmark for LACIE (Stengel-Eskin et al., 2024) and as out-of-distribution evaluation for SaySelf and ConfTuner. Its model-based judging protocol also lets truthfulness and informativeness be reported separately, which is useful precisely because hedging improves one at the expense of the other.

HaluEval (Li et al., 2023) provides generated and human-annotated hallucinated responses, and is used by Zhang et al. (2024b) as an out-of-domain refusal test. **FEVER** (Thorne et al., 2018) and **WiCE** (Kamoi et al., 2023) cast claim verification as multiple choice and enter the same multi-task refusal setting. **FalseQA** (Hu et al., 2023) contains questions built on false presuppositions, where the correct behaviour is to reject the premise rather than answer, and **SelfAware** (Yin et al., 2023) contains unanswerable questions paired with answerable ones, making it the closest thing the field has to a direct test of knowing what one does not know.

LAMA (Petroni et al., 2019) and **ParaRel** (Elazar et al., 2021) probe parametric knowledge with subject–relation–object facts phrased as cloze completions, with ParaRel adding multiple paraphrases of each relation. The cloze format is what makes them suitable for token-level interventions: Cohen et al. (2024) shift probability mass toward an [IDK] token at exactly the position where the object would be produced, and Zhang et al. (2024b) use ParaRel to partition an instruction dataset by what the base model already knows.

B.6 Ambiguous and Underspecified Questions

Most standard QA benchmarks assume that each question has a single intended answer, making disagreement among model generations easy to interpret as uncertainty or error. Ambiguous and underspecified questions break this assumption: several different answers may be equally valid because the prompt permits multiple interpretations. These datasets are therefore important for testing whether an uncertainty method can distinguish genuine ambiguity in the input from uncertainty caused by limitations in the model's own knowledge.

AmbigQA (Min et al., 2020) annotates ambiguous NQ questions with their distinct interpretations and the answer belonging to each. Hou et al. (2024) use it, together with an NLI counterpart they construct, to test whether input clarification ensembling recovers the aleatoric component of uncertainty. Tomov et al. (2026) go further and introduce **MAQA*** and **AmbigQA***, the first ambiguous QA benchmarks equipped

with ground-truth answer *distributions* derived from factual co-occurrence statistics in Wikipedia, which makes it possible to check whether an estimator recovers the true multiplicity of answers rather than merely producing a large number. This is the main gap in the current dataset landscape: almost every other benchmark assumes a single correct answer and therefore cannot distinguish ignorance from ambiguity by construction.

B.7 Long-Form Generation

Response-level scoring on short answers is known to overstate reliability, and a small set of benchmarks exists to test the long-form case.

Biography generation, introduced as the evaluation set of FactScore (Min et al., 2023) and used under the names **Bios**, **BioGen**, and **FactualBio**, prompts the model to write a biography of a named person, after which the paragraph is decomposed into atomic claims and each is verified. It is used by Farquhar et al. (2024), Zhang et al. (2024a), and Tjandra et al. (2024). **LongFact** (Wei et al., 2024) and **WildHallu** (Zhao et al., 2024) extend this to broader topic sets, the latter derived from real user conversations, and both are used by Zhang et al. (2024a) to show that macro-level calibration hides atomic-level miscalibration.

Long-context QA plays a different role. Zhang et al. (2024c) build their verbosity study on **Qasper** (Dasigi et al., 2021), questions over NLP papers; a composite of **HotpotQA**, **MuSiQue** (Trivedi et al., 2022), and **2WikiMultihopQA** (Ho et al., 2020) drawn from **LongBench** (Bai et al., 2023); **NarrativeQA** (Kočiský et al., 2018), questions over entire books and film transcripts; and short-answer subsets of NQ and MMLU. Here the context length is the independent variable: the gold answers are constrained to four words or fewer, so any additional words in the response are compensation rather than content.

B.8 Purpose-Built Calibration Suites

Unlike most datasets discussed above, which were originally developed for question answering or reasoning and later repurposed for uncertainty evaluation, a small number of benchmarks were designed specifically to test whether models recognize the limits of their own predictions.

CalibratedMath (Lin et al., 2022a) consists of synthetic arithmetic tasks constructed to evaluate whether a model's reported confidence matches its actual probability of being correct. The tasks vary systematically in difficulty and type, providing a controlled setting in which calibration can be studied independently of the broad factual and linguistic variation present in general-purpose QA benchmarks. Its task-level splits also make it possible to evaluate whether calibrated confidence transfers to related but previously unseen problem types.

SelfAware (Yin et al., 2023) instead focuses on whether models can recognize questions that they should not answer. It contains both answerable and unanswerable questions, allowing evaluation of whether a model can distinguish cases supported by its knowledge from those for which an appropriate response is to acknowledge uncertainty. Together, these datasets provide more direct tests of uncertainty awareness than conventional benchmarks, although they cover only two relatively narrow aspects of the problem: confidence calibration and recognition of unanswerability.

B.9 Limitations of the Current Dataset Landscape

Four properties of this collection constrain what the field can currently conclude. First, coverage is heavily skewed toward short-form English factoid QA, where a single answer string can be matched automatically. The settings in which models are actually deployed are long-form and multilingual. Second, almost all of these benchmarks presuppose one correct answer, so an estimator that conflates ambiguity with ignorance is never penalised for it, and Tomov et al. (2026) show that performance on TriviaQA-style data does not predict performance once that assumption is dropped. Third, the same handful of datasets is reused as both training and evaluation data across papers, and several benchmarks derived from Wikipedia are near-certain to appear in the pretraining corpora of the models being evaluated, which inflates accuracy and shrinks the low-confidence region that uncertainty methods are meant to detect. Fourth, correctness itself is increasingly

judged by another language model, in TruthfulQA, in the long-form biography settings, and in the sentence-length variants of the semantic entropy evaluations, which introduces a second, uncalibrated source of error into the very labels used to score calibration.

Table 3: Datasets used across the reviewed uncertainty-awareness literature.

Dataset	**Task format**	**Answer length**	**Representative use in this survey**
TriviaQA	Closed-book trivia QA	Phrase	Semantic entropy, KLE, LACIE, IDK-tuning, ConfTuner
NQ / NQ-Open	Real search queries over Wikipedia	Phrase	Probabilistic vs. verbalized confidence, faithfulness
SQuAD	Extractive reading comprehension	Phrase/sentence	Semantic entropy and kernel variants
PopQA	Entity facts with popularity labels	Phrase	Rare vs. frequent knowledge, [IDK] token
CoQA	Conversational QA with coreference	Free-form	Black-box graph estimators, semantic uncertainty
SciQ	Science exam questions	Phrase	Correctness-predictor training, OOD calibration
BioASQ	Expert biomedical QA	Phrase/sentence	Out-of-domain probe for semantic estimators
MedQA	Medical licensing exam MCQ	Option	Confidence-token routing
GSM8K	Grade-school math word problems	Number	Confidence elicitation, consistency calibration
SVAMP	Adversarial math word problems	Number	Confidence elicitation, semantic entropy
ASDiv / MultiArith	Math word problems	Number	Sample-consistency calibration
StrategyQA	Implicit multi-hop yes/no	Binary	Confidence elicitation, consistency calibration
BIG-bench subsets	Symbolic and commonsense reasoning	Phrase	Confidence elicitation across task types
SayCan / CLUTRR	Planning; relational inference	Structured	Sample-consistency calibration
HotpotQA	Multi-hop QA over Wikipedia	Phrase	SaySelf, ConfTuner, R-Tuning
MMLU	57-subject MCQ	Option	Refusal tuning, routing, knowledge retention
ARC / OpenBookQA / CommonsenseQA	Science and commonsense MCQ	Option	Correctness-predictor training, retention checks
HellaSwag / WinoGrande	Commonsense completion MCQ	Option	Retention checks after uncertainty tuning
TruthfulQA	Misconception-prone questions	Sentence	LACIE transfer, OOD calibration
HaluEval	Annotated hallucinated responses	Free-form	OOD refusal evaluation
FEVER / WiCE	Claim verification as MCQ	Label	Multi-task refusal tuning
FalseQA	False-presupposition questions	Free-form	Refusal-aware tuning
SelfAware	Answerable vs. unanswerable questions	Free-form	Direct test of known unknowns
LAMA / ParaRel	Cloze-style relational facts	Entity	[IDK] token, knowledge-boundary partitioning
AmbigQA	Ambiguous questions with interpretations	Phrase	Uncertainty decomposition
MAQA* / AmbigQA*	Ambiguous QA with answer distributions	Phrase	UQ evaluation under ambiguity
FactScore / Bios	Biography generation	Paragraph	Atomic calibration, long-form abstention
LongFact / WildHallu	Long-form factuality prompts	Paragraph	Atomic calibration
Qasper / LongBench / NarrativeQA	Long-context QA	Short gold answers	Verbosity compensation
CalibratedMath	Synthetic arithmetic with confidence targets	Number	Supervised verbalized calibration