eSystems Engineering Society EST. 2007

International Journal of Data Science and Advanced Analytics

Journal homepage: www.ijdsaa.com

Uncertainty, Disagreement and Information Fusion in Modern AI Systems: A Comprehensive Survey

Mohammad Kasra Sartaee[1], Mary Qasim Al-Shihani[2]
[1]University College London, UK, [2]Independent Researcher, UAE, Dubai

[1] Mokasoftuk@gmail.com, [2]Shihanimary@gmail.com

Abstract- The reliability of modern artificial intelligence (AI) systems critically depends on their ability to quantify uncertainty, interpret disagreement and fuse information from multiple models or data sources. While classical machine learning emphasised probability calibration for discriminative classifiers, contemporary AI systems increasingly operate in open-ended, generative and multimodal settings where uncertainty is semantic, subjective and decision-dependent. This survey provides a comprehensive review of uncertainty estimation, disagreement modelling, information fusion and selective prediction in modern AI systems. We formalise aleatoric and epistemic uncertainty and review foundational evaluation metrics including negative log-likelihood, proper scoring rules and calibration error. We then survey uncertainty estimation methods spanning Bayesian neural networks, deep ensembles, calibration techniques, consistency-based approaches, conformal prediction, evidential deep learning and semantic uncertainty measures for generative models. Disagreement is considered as an informative signal rather than noise, covering ensemble disagreement, human annotation variability and multi-agent systems. We review uncertainty-aware information fusion techniques for multimodal and multi-model systems, including classical Bayesian fusion, attention-based gating mechanisms and modality-dropout strategies. Throughout the survey, we highlight empirical trade-offs between computational cost and reliability, provide benchmark-anchored comparisons, discuss evaluation practices, and identify open challenges related to scalability, distribution shift, hallucination detection, and decision-theoretic integration. The result is a unified reference for researchers and practitioners designing AI systems that are not only accurate but reliably aware of their own limitations.

Keywords- Uncertainty Quantification; Disagreement Modelling; Information Fusion; Calibration; Selective Prediction; Conformal Prediction; Evidential Deep Learning; Generative Models; Reliable AI.

1. Introduction

Assessing the reliability of artificial intelligence (AI) systems has long been a central concern in machine learning (ML). Traditionally, reliability was closely associated with calibration, defined as the alignment between a model's predicted confidence and its empirical accuracy [1]. This paradigm has been extensively studied for discriminative classifiers and remains foundational for uncertainty quantification. Recent advances in large language models (LLMs), generative vision systems and multimodal agents have fundamentally altered the landscape of AI reliability.

In the aforementioned systems, outputs are drawn from open-ended spaces where correctness is semantic rather than categorical [2]. As a result, uncertainty arises not only from statistical noise, but also from lack of clear specifications, ambiguity, subjectivity and model limitations. This shift challenges classical notions of reliability based solely on scalar confidence scores or

https://doi.org/10.69511/ijdsaa.v7i2.327
Received 10 June 2025; Received in revised form 20 August 2025; Accepted 5 October 2025.
Available online 10 October 2025.

token-level entropy, motivating uncertainty representations that operate over semantic meaning and agreement rather than surface form.

In parallel, modern AI systems increasingly rely on ensembles, pipelines of specialised models, and multimodal architectures. Disagreement among models, agents and/or human annotators is no longer merely a nuisance to be averaged away, but a valuable signal reflecting epistemic uncertainty or intrinsic task ambiguity [3]. Similarly, robust decision-making requires principled mechanisms for fusing information from multiple uncertain sources [4].

The present review addresses four interconnected pillars of modern AI reliability: uncertainty estimation, disagreement modelling, information fusion and selective prediction. We synthesise theoretical foundations, empirical methods and evaluation practices, emphasising trade-offs between rigour, scalability and real-world applicability. Our contribution is distinct from prior reviews in several ways. Gawlikowski et al. [5] provided a comprehensive taxonomy of uncertainty methods but did not address disagreement modelling or information fusion as separate pillars. Abdar et al. [6] explored uncertainty quantification techniques with an emphasis on applications but gave limited attention to generative models and semantic uncertainty. Hüllermeier and Waegeman [7] offered a rigorous conceptual treatment of aleatoric and epistemic uncertainty but did not cover practical methods for large-scale models. To complement these studies, the present review contributes by integrating all four pillars being uncertainty, disagreement, fusion and selective prediction. This and covers the recent shift toward semantic uncertainty in LLMs, providing a unified reference that bridges classical foundations with contemporary developments.

2. Methodology

A systematic literature review was conducted considering the PRISMA reporting guidelines [46]. The review was conducted using five search engines: Google Scholar, Semantic Scholar, arXiv, IEEE Xplore and ACM Digital Library using the following keyword groups: (a) “uncertainty estimation” OR “uncertainty quantification” OR “predictive uncertainty” combined with “deep learning” OR “neural networks” OR “large language models”; (b) “disagreement” OR “annotator disagreement” OR “ensemble disagreement” combined with “machine learning”; (c) “information fusion” OR “multimodal fusion” combined with “uncertainty”; (d) “selective prediction” OR “abstention”. The systematic search covered publications from January 2015 to March 2026; seminal earlier works (e.g., classical decision theory, Bayesian filtering and evidence theory) were additionally included as background references through manual citation chaining and lie outside the systematic corpus. Where required, reference lists of included studies were manually searched for additional sources.

Inclusion criteria were studies that: (i) presented novel uncertainty estimation, disagreement modelling, uncertainty-aware fusion methods, or selective prediction; (ii) published in peer-reviewed venues, or available as widely cited and influential preprints; (iii) addressed deep learning (DL) systems specifically. We excluded from the systematic corpus papers focused exclusively on classical statistical methods without neural network components (classical foundations are nevertheless summarised as background in Section 6.1) and application-specific studies without methodological contributions.

The final review corpus comprised 45 papers. Prior to data extraction, each paper was appraised for methodological transparency using criteria adapted from the machine-learning reproducibility checklist [47] (clarity of experimental protocol, availability of code and data, and completeness of reported evaluation details). Lower-scoring studies were retained but are interpreted with caution.

The review concentrates on uncertainty quantification, disagreement modelling, information fusion and selective prediction for discriminative and generative models. Adjacent areas, uncertainty in reinforcement learning and active learning, Bayesian optimisation, causal uncertainty, and uncertainty in retrieval-augmented generation, are out of scope, except where individual results bear directly on the four pillars.

3. Foundations of Uncertainty in AI

3.1. Aleatoric and Epistemic Uncertainty

Uncertainty in ML predictions is commonly decomposed into two components [8]. Aleatoric uncertainty refers to irreducible variability inherent in the data-generating process, such as sensor noise or intrinsic ambiguity in tasks with multiple valid outputs. This form of uncertainty cannot be eliminated even with infinite data. Epistemic uncertainty reflects uncertainty due to limited model knowledge, arising from insufficient data, model misspecification or incomplete coverage of the input space [9]. In principle, epistemic uncertainty is reducible through additional diverse data or improved modelling assumptions.

Kendall and Gal [10] provided a practical framework for decomposing these two forms within Bayesian DL for computer vision, demonstrating that aleatoric uncertainty can be captured through learned observation noise parameters while epistemic uncertainty is estimated via approximate posterior inference over model weights. Hüllermeier and Waegeman [7] further formalised these concepts, noting that the distinction is not always clean in practice: in modern DL systems, especially generative models, high predictive entropy may reflect either genuine ambiguity or lack of model competence, motivating explicit modelling approaches [2].

3.2. Evaluation Metrics

Reliable uncertainty estimation requires evaluation metrics that connect probabilistic outputs to empirical outcomes. Proper scoring rules assess the quality of predictive distributions by rewarding calibrated, sharp probabilities [11]. Two widely used proper scoring rules are negative log-likelihood (NLL) and the Brier score. Given a dataset $\{(x_i, y_i)\}$from i = 1 to n and predicted class probabilities p(y|x), the NLL is:

$$\mathcal{L}_{\mathcal{NLL}} = -\frac{1}{n}\sum_{i=1}^{n}\log p_\theta(y_i \mid x_i) \quad (1)$$

For multiclass classification with one-hot target vector $y_i \in \{0,1\}^K$ and predicted probability vector $\hat{p}_i \in [0,1]^K$, the Brier score is:

$$Brier = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\left(\widehat{p_{i,k}} - y_{i,k}\right)^2 \quad (2)$$

Calibration metrics measure the discrepancy between predicted confidence and empirical accuracy. Expected Calibration Error (ECE) partitions predictions into M confidence bins $\{B_m\}$ and computes:

$$ECE = \sum_{m=1}^{M}\frac{|B_m|}{n}|acc(B_m) - conf(B_m)| \quad (3)$$

where $\mathrm{acc}(B_m)$ is the average accuracy of samples in bin B_m and $\mathrm{conf}(B_m)$ is their average predicted confidence [1].

Beyond calibration, uncertainty estimates are frequently evaluated through downstream tasks such as error detection, selective prediction, and out-of-distribution (OOD) detection, using metrics such as the area under the receiver operating characteristic curve (AUROC) [12]. Each metric captures different aspects of uncertainty quality and no single measure is sufficient in isolation. Two caveats apply: ECE is sensitive to the binning scheme and is not a proper scoring rule, so it should be read alongside NLL or the Brier score; and AUROC, while threshold-independent, can appear optimistic under strong class imbalance.

4. Uncertainty Estimation Methods

4.1. Bayesian Neural Networks

Bayesian neural networks (BNNs) model epistemic uncertainty by maintaining a posterior distribution over model parameters and marginalising predictions over this distribution [9]. The Bayesian approach to neural networks (NNs) was pioneered by MacKay [39], who developed practical frameworks for Bayesian inference in adaptive models. Exact posterior inference is intractable for NNs, so approximate techniques are employed. Variational inference methods, such as Bayes by Backprop [13], learn a parametric approximation to the posterior by minimising the Kullback-Leibler divergence. Monte Carlo dropout [14] recasts standard dropout as approximate variational inference, enabling uncertainty estimation from any dropout-trained network without architectural modification. Laplace approximations fit a Gaussian to the posterior around a mode of the loss landscape [15], [51]. A related line of work uses mixture density networks [38] to model conditional output distributions, capturing aleatoric uncertainty through explicit distributional outputs.

While theoretically principled, BNN approaches face scalability challenges. Variational methods introduce additional parameters and often underestimate posterior variance due to mean-field assumptions. Monte Carlo (MC) dropout, though computationally cheaper, provides coarse approximations that may not capture multimodal posteriors. Empirically, the performance of BNNs depends heavily on the choice of prior, approximate inference method and network architecture. In addition, BNNs frequently underperform deep ensembles on calibration and OOD detection benchmarks [12].

4.2. Deep Ensembles

Deep ensembles estimate uncertainty by aggregating predictions from independently trained models [16]. Given L independently trained models with predictive distributions $\{p_{\theta_l}(y \mid x)\}$ for $l = 1, \ldots, L$, the ensemble predictive distribution is the mixture:

$$p_{ens}(y|x) = \frac{1}{L}\sum_{l=1}^{L} p_{\theta_l}(y|x) \quad (4)$$

Epistemic uncertainty can be estimated via the variance of predicted probabilities across ensemble members. For a K-class classification problem, the ensemble disagreement is measured as:

$$Disagree(x) = \frac{1}{K}\sum_{k=1}^{K} Var_l\left[p_{\theta_l}(y = k|x)\right] \quad 5$$

where $Var_l[\cdot]$ denotes the variance across the L ensemble members for each class k, yielding a scalar measure of per-class prediction spread averaged over all classes.

On the benchmark study of Ovadia et al. [12], deep ensembles of five independently trained ResNets consistently outperformed single-model baselines, MC dropout, and variational inference methods across multiple dataset-shift scenarios (CIFAR-10-C, ImageNet-C), achieving the lowest ECE and highest AUROC for misclassification detection. Lakshminarayanan et al. [16] similarly demonstrated that ensembles achieve well-calibrated predictive uncertainty on regression and classification tasks without requiring explicit Bayesian treatment.

However, ensembles are computationally expensive, requiring L times the training and inference cost of a single model. Recent work on efficient ensemble methods, such as BatchEnsemble [17] which shares most parameters across members and only maintains per-member rank-one perturbations, addresses some of these costs. Nevertheless, a growing body of work has identified epistemic uncertainty collapse in large or over-parameterised models, where ensemble diversity diminishes as model capacity increases [53], limiting the reliability of variance-based uncertainty estimates.

4.3. Calibration and Post-hoc Methods

Post-hoc calibration techniques adjust predicted confidence without modifying the underlying model. Temperature scaling [1] rescales logits z(x) by a learned scalar T > 0:

$$\widehat{p_T}(y|x) = softmax\left(\frac{z(x)}{T}\right)_y \quad 6$$

Guo et al. [1] demonstrated that modern NNs are systematically overconfident and that temperature scaling, learned on a held-out validation set, substantially reduces ECE on CIFAR-10, CIFAR-100, and ImageNet. However, a single global temperature parameter cannot capture heterogeneous miscalibration across different input regions. Under distribution shift, temperature scaling degrades significantly. Ovadia et al. [12] showed that post-hoc calibrated models exhibit worse ECE than uncalibrated ensembles when evaluated on corrupted versions of standard benchmarks. More recent approaches, including focal loss calibration and input-dependent temperature functions, partially address this limitation but remain less robust than ensemble-based methods under severe shift.

4.4. Consistency-Based Methods

Consistency-based approaches estimate uncertainty by measuring variability across multiple stochastic generations. Self-consistency [18] assumes that reliable reasoning produces stable outputs: given a question, multiple reasoning chains are sampled, and the most frequent answer is selected, with answer frequency serving as a confidence proxy. SelfCheckGPT [19] extends this idea to hallucination detection by checking whether factual claims in a generated response are consistent across independent samples.

While effective in some settings, these methods incur significant computational cost requiring O(J) forward passes for J samples. Furthermore, the improvements these methods provide over simpler baselines are not always commensurate with that cost. Kadavath et al. [20] found that self-evaluation and sampling-based measures correlate with answer correctness on question-answering benchmarks, but that gains over simple token-level probability baselines can be modest relative to the multiple-forward-pass inference cost.

4.5. Conformal Prediction

Conformal prediction provides a distribution-free framework for constructing prediction sets with guaranteed coverage, a framework originating with Vovk et al. [48] and brought to deep learning practice by Angelopoulos and Bates [21]. Given a calibration dataset, conformal methods compute non-conformity scores and use quantiles of these scores to construct prediction sets that contain the true label with probability at least $1 - \alpha$, under the sole assumption of exchangeability. This guarantee is marginal: it holds on to average over exchangeable draws of the calibration and test data, and does not by itself imply conditional coverage for individual inputs or population subgroups without stronger assumptions or modified procedures. Romano et al. [22] introduced conformalised quantile regression, which combines quantile regression with conformal calibration to produce adaptive prediction intervals for regression tasks. Angelopoulos et al. [23] proposed the Regularised Adaptive Prediction Sets (RAPS) method, which produces smaller prediction sets for image classification by penalising set size during calibration.

The appeal of conformal prediction lies in its finite-sample validity guarantee, which holds regardless of the underlying model or data distribution. However, the

practical utility depends on the predictive quality of the underlying model and the informativeness of the non-conformity score: a weak model yields valid but uninformatively large prediction sets. Furthermore, standard conformal methods assume exchangeability, which is violated under distribution shift. Recent extensions address this limitation through weighted conformal inference and online conformal prediction, but these remain active areas of research.

4.6. Evidential Deep Learning

Evidential DL methods place Dirichlet (for classification) or Normal-Inverse-Gamma (for regression) priors over the output distribution and train the network to predict the parameters of this prior [24]. Rather than outputting point predictions or class probabilities, the network outputs evidence parameters from which both aleatoric and epistemic uncertainty can be derived in a single forward pass. Amini et al. [25] extended this framework to regression tasks with Deep Evidential Regression, predicting the four parameters of a Normal-Inverse-Gamma distribution.

A key advantage of evidential methods is computational efficiency: unlike ensembles or BNNs, they require only a single forward pass at inference time. Charpentier et al. [26] introduced Posterior Networks, which use normalising flows to estimate density-based pseudo-counts and obtain Dirichlet posteriors without explicitly training on OOD data. This was subsequently extended to exponential family distributions with the Natural Posterior Network [42], enabling evidential uncertainty for a broader class of output distributions. However, evidential methods face known limitations related to the choice of regularisation and degradation in the quality of their uncertainty estimates. Ulmer et al. [27] provide a comprehensive analysis of these trade-offs in their survey of prior and posterior network methods.

4.7. Semantic Uncertainty for Generative Models

For generative models, token-level entropy often fails to capture semantic equivalence. For instance, two responses being “The capital of France is Paris” and “Paris is France’s capital city” are semantically identical but differ at the token level. Semantic uncertainty methods [2] address this by clustering generated outputs based on meaning rather than surface form.

Semantic entropy groups J sampled outputs $\{y^{(j)}\}$ into semantic clusters $C = \{c_1, \dots, c_M\}$ using a bidirectional entailment classifier. The estimated cluster probability is:

$$\hat{p}(c_m|x) = \sum_{j:y^{(j)} \in c_m} \hat{p}\left(y^{(j)}|x\right) \quad (7)$$

where $\hat{p}\left(y^{(j)} \mid x\right)$ is the length-normalised sequence probability of the j-th sampled generation; cluster probabilities are renormalised over the M clusters observed in the sample, and semantic entropy is evaluated as a Monte Carlo estimate over these clusters [2]. Semantic entropy is then defined as:

$$H_{sem}(x) = -\sum_{m=1}^{M} \hat{p}(c_m|x) \log \hat{p}(c_m|x) \quad (8)$$

This collapses probability mass over linguistically distinct but semantically equivalent outputs [2]. A baseline comparison is token-level predictive entropy. For a vocabulary V and next-token distribution $p_\theta(\cdot \mid x, y_{<t})$, entropy at position t is:

$$H_{token}(x,t) = -\sum_{v \in V} p_\theta(v|x, y_{<t}) \log p_\theta(v|x, y_{<t}) \quad (9)$$

where the summation is over all tokens v in the vocabulary V at position t, conditioned on the input x and all preceding tokens $y_{<t}$. Total sequence entropy can be obtained by averaging across positions. However, in natural language generation, high H_{token} may reflect surface variation (e.g., synonym choice) rather than genuine uncertainty about meaning.

Kuhn et al. [2] demonstrated that semantic entropy outperforms token-level predictive entropy and related sampling-based baselines for detecting incorrect answers on the TriviaQA and CoQA benchmarks, with consistent AUROC improvements across OPT models ranging from 2.7B to 30B parameters. Farquhar et al. [28] extended this work to hallucination detection on a broader benchmark suite (including SQuAD, BioASQ and Natural Questions), showing that semantic entropy reliably detects confabulated claims across multiple LLM families. Complementary approaches explore verbalised uncertainty, where LLMs are trained to express confidence in natural language [40], and prompt-based confidence elicitation, where calibration is assessed by asking models to self-report certainty [41]. Kernel-based

generalisations capture graded semantic similarity and provide fine-grained uncertainty estimates, though at increased computational expense.

4.8. Out-of-Distribution Detection

Out-of-distribution detection is closely related to uncertainty estimation, as models should express high uncertainty on inputs far from the training distribution. Hendrycks and Gimpel [29] established the maximum softmax probability (MSP) baseline, which flags inputs where the maximum predicted probability falls below a threshold. Liang et al. [30] improved upon this with the ODIN detector, combining temperature scaling with input perturbations. Liu et al. [31] proposed energy-based OOD detection, replacing softmax confidence with the energy function $E(x) = -\log \sum_k \exp\left(f_k(x)\right)$, demonstrating consistent improvements over MSP and ODIN across standard OOD benchmarks (CIFAR-10 vs. SVHN, CIFAR-100 vs. LSUN). While OOD detection methods are related to but distinct from uncertainty estimation, they are increasingly integrated into unified reliability frameworks.

5. Disagreement as an Informative Signal

Disagreement has traditionally been treated as noise to be eliminated through majority voting or averaging. Recent work fundamentally reframes disagreement as an informative signal that reflects either epistemic uncertainty (reducible through more data or better models) or intrinsic task ambiguity (irreducible aleatoric uncertainty) [3]. This section examines disagreement across three settings: ensemble models, human annotations and multi-agent systems.

5.1. Ensemble Disagreement

In ensemble models, disagreement among member predictions provides a direct estimate of epistemic uncertainty. Jiang et al. [3] formalised this through the Generalisation Disagreement Equality (GDE), which establishes that the expected disagreement rate between two networks trained with independent runs of stochastic gradient descent equals the expected test error, provided the induced ensemble is well calibrated in the class-aggregated sense. This theoretical result enables estimating model accuracy on unlabelled data purely from disagreement, without requiring ground-truth labels.

The practical implications are significant. Under distribution shift, where validation accuracy is unavailable, ensemble disagreement can serve as a proxy for error rate, enabling automatic detection of deployment scenarios where model reliability has degraded. An important caveat is that the calibration condition underpinning the GDE itself deteriorates as disagreement grows — precisely the regime in which the error–disagreement coupling would be most valuable — so disagreement-based monitoring must be applied with care under shift [52]. Empirically, Ovadia et al. [12] showed on CIFAR-10-C and ImageNet-C that the quality of predictive uncertainty degrades with increasing corruption severity, with ensembles degrading most gracefully among the methods compared.

However, the GDE assumes that ensemble members explore genuinely different hypotheses. When models converge to similar solutions, as is the case for large over-parameterised architectures trained on the same data, disagreement collapses and the epistemic uncertainty signal is lost. This phenomenon, termed uncertainty collapse, has been documented across vision and language tasks [53] and represents a fundamental limitation of disagreement-based approaches for very large models.

5.2. Human Annotation Disagreement

In subjective tasks such as sentiment analysis, toxicity detection and medical diagnosis, annotator disagreement is not labelling error but a genuine reflection of aleatoric uncertainty arising from task ambiguity [56]. Traditionally, ML pipelines resolve disagreement through majority voting, discarding the distributional information in label distributions. Recent work argues that this practice is epistemologically problematic: it imposes a false consensus and prevents models from learning that certain inputs are inherently ambiguous.

Learning with disagreement approaches model the full distribution of human judgements rather than collapsing to a single label. Soft-label training, which uses the empirical distribution of annotator responses as the target, has been shown to improve calibration and robustness in image classification [54] and on subjective NLP tasks [55]. Multi-annotator models explicitly parameterise annotator-specific biases, separating genuine ambiguity from individual annotator noise [55]. These approaches are particularly relevant for safety-critical applications where overconfident predictions on ambiguous inputs carry significant risk.

5.3. Disagreement in Multi-Agent Systems

In multi-agent AI systems — including LLM debate frameworks [57], mixture-of-experts architectures [58], and multi-model pipelines — disagreement among agents provides a natural uncertainty signal. Aggregation mechanisms such as voting, weighted fusion, or deliberative debate are used to combine diverse agent

outputs. However, recent theoretical and empirical analyses demonstrate that interaction alone does not guarantee correctness gains.

When agents share training data, architectural biases or are derived from the same foundation model, they may exhibit correlated errors that simple aggregation cannot correct. In LLM debate settings, models may converge on plausible but incorrect consensus answers through mutual reinforcement rather than genuine reasoning [57]. Effective multi-agent uncertainty estimation therefore requires explicit mechanisms for calibration, diversity enforcement and weighting by estimated competence. Disagreement must be interpreted in conjunction with individual agent uncertainty estimates to avoid reinforcing shared biases or systematic errors.

6. Information Fusion

Information fusion integrates multiple uncertain sources into a coherent decision. As AI systems increasingly combine heterogeneous data modalities, model outputs and sensor streams, principled fusion mechanisms that account for source-specific uncertainty become essential.

6.1. Classical Bayesian Fusion

The foundations of uncertainty-aware fusion lie in Bayesian inference and Dempster-Shafer theory [32]. The Kalman filter [33] provides the canonical example of optimal Bayesian fusion for linear Gaussian systems, weighting measurements by the inverse of their uncertainty (covariance). This principle generalises: given N independent sources with estimates $\{\mu_i\}$ and uncertainties $\{\sigma_i^2\}$, the minimum-variance fused estimate weights each source inversely proportional to its variance. Bayesian model averaging extends this to combining predictions from multiple models, weighting each by its posterior probability.

While optimal under correct model assumptions, classical Bayesian fusion requires accurate uncertainty estimates from each source and explicit specification of the joint model. In practice, both conditions are often violated in deep learning systems, motivating learned fusion approaches.

6.2. Uncertainty-Aware Multimodal Fusion

Modern multimodal systems extend classical fusion by learning mechanisms that dynamically adjust modality contributions based on estimated reliability. Gao et al. [4] proposed embracing unimodal aleatoric uncertainty for robust fusion: the noise inherent in each modality is explicitly quantified and used within uncertainty-aware contrastive learning to produce stable unimodal embeddings prior to fusion, improving robustness to noisy modalities on multimodal classification benchmarks. A complementary line of work performs quality-aware dynamic fusion, weighting each modality at inference time according to its estimated quality or uncertainty, with provable robustness guarantees for low-quality multimodal data [50].

Attention-based gating mechanisms provide another approach to uncertainty-aware fusion. Cross-modal attention architectures learn to attend selectively to reliable modalities, effectively implementing a soft version of the Bayesian weighting principle. Modality-dropout training, where random modalities are zeroed during training, forces the model to handle missing or unreliable inputs, improving robustness at inference time when modality quality varies.

6.3. Multi-Model Fusion and Aggregation

Beyond multimodal fusion, uncertainty-aware aggregation is essential when combining outputs from multiple models in ensemble or pipeline architectures. Weighted prediction aggregation schemes assign weights to model outputs based on estimated confidence or historical performance. Mixture-of-experts architectures [58] use gating networks to route inputs to specialised sub-models, with the gating mechanism implicitly performing uncertainty-aware selection. Recent work has explored integrating explicit uncertainty estimates into the gating function, improving routing decisions for out-of-distribution inputs.

A critical challenge in multi-model fusion is calibration consistency: if individual models are mis-calibrated in systematically different ways, naive fusion can amplify rather than reduce errors. This underscores the importance of the uncertainty–disagreement–fusion pipeline: reliable fusion requires reliable uncertainty estimates, which in turn benefit from meaningful disagreement signals.

7. Selective Prediction and Abstention

Selective prediction allows models to abstain when uncertainty is high, trading coverage (the fraction of inputs on which a prediction is made) against selective risk (the error rate on retained inputs). Performance is therefore best evaluated with risk–coverage curves or the area under the risk–coverage curve, rather than accuracy at a single operating point. This section connects abstention directly to the preceding pillars: the quality of selective prediction depends critically on the uncertainty estimates

reviewed in Section 4, disagreement signals from Section 5 can trigger abstention, and fusion architectures from Section 6 can incorporate abstention as a routing decision.

Classical decision-theoretic results define optimal rejection rules under known costs [34]. The Chow reject option establishes that optimal abstention occurs when the posterior probability of the predicted class falls below a cost-dependent threshold. Modern approaches integrate abstention into training: SelectiveNet [35] jointly optimises a prediction head, a selection head, and a coverage-accuracy trade-off via a risk-coverage objective.

The effectiveness of selective prediction depends critically on uncertainty quality. Ovadia et al. [12] showed that under distribution shift on corrupted image benchmarks, the accuracy of retained predictions degrades far less for deep ensembles than for single-model softmax confidence as the confidence threshold is varied, so ensemble-based abstention retains substantially higher accuracy at matched coverage. This result directly links the quality of uncertainty estimation methods (Section 4) to downstream decision-making performance.

8. Comparative Analysis of Methods

This section provides a structured comparison of the methods reviewed above, anchoring qualitative assessments to empirical evidence where available.

Table 1 compares key properties across method families. Table 2 summarises representative empirical findings from the underlying studies; because these studies differ in architectures, model scales and evaluation protocols, findings are reported qualitatively rather than as directly comparable point estimates.

Table 1 Comparison of uncertainty and disagreement estimation approaches. [1, 2, 12, 16, 24].

Method	Aleatoric/ Epistemic	OOD Detect	Key Limitation
Token Entropy [a]	Neither explicitly	Weak	Ignores semantic equivalence
MC Dropout [14][a]	Epistemic only	Moderate	Coarse posterior approx.
Deep Ensembles [16][a]	Both (via entropy decomp.)	Strong	L× compute; collapse risk
Temp. Scaling [1][a]	Neither	None	Fails under shift
Conformal Pred. [21][a]	Neither (coverage)	N/A	Set size depends on model
Evidential DL [24][a]	Both	Variable	Regularisation-sensitive
Semantic Entropy [2][b]	Semantic-cluster uncertainty (no decomp.)	Strong	Requires NLI classifier
SelfCheck GPT [19][c]	Consistency signal (no decomp.)	Moderate	O(J) sampling cost; misses consistent confabulations
Unc.-Aware Fusion [4][c]	Aleatoric per modality	Moderate	Requires per-source uncertainty

a: does not capture semantic equivalence; b: captures semantic equivalence; c: partially captures semantic equivalence. OOD: out-of-distribution; NLI: natural language inference; decomp.: decomposition of predictive entropy into expected entropy (aleatoric) and mutual information (epistemic).

Table 2 Representative empirical findings for uncertainty estimation methods, as reported in the cited studies.

Method	Benchmark	Reported finding	Cost	Source
Deep ensemble (L=5) vs MC dropout, VI, single model	CIFAR-10(-C), ImageNet(-C)	Best calibration (ECE, Brier) among compared methods; uncertainty quality degrades most gracefully as corruption severity increases	L×	Ovadia et al. [12]
MC dropout	CIFAR-10(-C)	Improves over a single model but is consistently outperformed by deep ensembles, especially under shift	T×	Ovadia et al. [12]
Temperature scaling	CIFAR-10/100, ImageNet (i.i.d.)	Substantially reduces ECE of modern networks post hoc on held-out i.i.d. data	~1×	Guo et al. [1]
Temperature scaling	CIFAR-10-C (shift)	Calibration degrades markedly under corruption relative to ensembles	~1×	Ovadia et al. [12]

Semantic entropy vs token-level entropy	TriviaQA, CoQA (OPT 2.7B–30B)	Consistently higher AUROC for detecting incorrect answers across model sizes	J×	Kuhn et al. [2]
Semantic entropy	QA suite incl. SQuAD, BioASQ, Natural Questions	Reliably detects confabulations across multiple LLM families	J×	Farquhar et al. [28]

AUROC: area under the receiver operating characteristic curve; ECE: expected calibration error; Inf. cost: inference cost relative to a single forward pass; L: ensemble size; T: number of dropout samples; J: number of sampled generations; VI: variational inference. Entries summarise findings as reported in the cited studies; values from different papers are not directly comparable.

Several patterns emerge from this comparison. First, deep ensembles remain the most robust general-purpose method for uncertainty estimation in classification, achieving the best calibration and OOD detection at the cost of L× compute. Second, post-hoc calibration (temperature scaling) is effective under i.i.d. conditions but degrades substantially under distribution shift, while ensemble-based methods degrade more gracefully. Third, for generative models, semantic uncertainty methods represent a qualitative advance over token-level approaches, with semantic entropy consistently outperforming token-level entropy for incorrect-answer detection across model sizes [2], demonstrating the importance of accounting for meaning equivalence. Fourth, the cost–performance trade-off is not automatic: sampling-based baselines operating at a computational budget comparable to semantic entropy do not close the gap to it [2], suggesting that the clustering step captures genuinely different information rather than simply benefiting from more samples.

Several limitations of these comparisons should be noted. Benchmark results are drawn from different papers with varying experimental configurations (model architectures, hyperparameters, data splits), and direct numeric comparison across papers should be interpreted cautiously. Furthermore, classification-oriented benchmarks (CIFAR-10, ImageNet) and LLM-oriented benchmarks (TriviaQA) test fundamentally different aspects of uncertainty quality and should not be naively ranked on the same scale.

9. Open Challenges

Despite significant progress, several challenges remain across all four pillars of this survey.

9.1. Scalability. Scalable uncertainty estimation for large generative models is unresolved. Ensemble and multi-sample methods require inference costs proportional to the number of members or samples, which may be prohibitive for models with billions of parameters. Single-forward-pass methods (evidential deep learning, spectral-normalised Gaussian processes [36]) offer computational advantages but have not yet been validated at the scale of frontier LLMs. Hybrid approaches such as deep kernel learning [43] combine neural feature extractors with Gaussian process output layers to balance expressiveness and uncertainty quality, but scalability to very deep architectures remains limited even with variational sparse approximations [45].

9.2. Distribution Shift. Many uncertainty methods degrade under distribution shift [12]. While conformal prediction offers finite sample guarantees under exchangeability, the assumption itself is violated in non-stationary deployment environments. Developing uncertainty methods that are robust under shift remains a critical open problem.

9.3. Cost–Utility Trade-offs. A growing body of empirical evidence highlights that multi-sample, ensemble-based, or clustering-driven approaches often incur substantial computational overhead while yielding marginal improvements over simpler baselines in downstream metrics such as AUROC. The practical deployment of advanced uncertainty methods in latency- or resource-constrained settings requires principled frameworks for deciding when expensive methods justify their cost.

9.4. Hallucination Detection. Detecting hallucinations in LLMs is among the most practically urgent applications of uncertainty estimation. While semantic entropy [2] and self-consistency approaches [18] show promise, they require multiple forward passes and cannot detect hallucinations that are consistently reproduced across samples (systematic confabulations). Efficient, single-pass hallucination detection remains an open problem.

9.5. Decision-Theoretic Integration. Uncertainty, disagreement, and fusion are ultimately in service of decision-making. Felicioni et al. [44] highlight the importance of integrating uncertainty estimates into downstream decision processes for LLMs. Tighter integration with decision theory, including cost-sensitive abstention, optimal delegation between models, and risk-aware planning, is needed to translate improved uncertainty estimates into improved outcomes.

9.6. Coverage Limitations of This Survey. We acknowledge several important areas that fall outside our scope. Gaussian processes, while foundational to uncertainty estimation, receive limited treatment here; the reader is referred to Rasmussen and Williams [37] for a comprehensive treatment. Normalising flows for density estimation and uncertainty are not covered. Our treatment of OOD detection is necessarily brief; for a dedicated survey, see Yang et al. [49]. We have also not covered uncertainty in reinforcement learning or federated learning settings.

10. Conclusion

Modern AI systems operate in environments where uncertainty is unavoidable and consequential. Reliable AI therefore requires principled uncertainty estimation, meaningful interpretation of disagreement, uncertainty-aware information fusion and selective prediction. This review covered these interconnected pillars, covering classical foundations (Bayesian methods, calibration, proper scoring rules), contemporary approaches (deep ensembles, conformal prediction, evidential deep learning, semantic uncertainty), and their interactions (disagreement-driven monitoring, uncertainty-weighted fusion, selective prediction).

The empirical evidence surveyed here supports several conclusions. Deep ensembles remain the most robust general-purpose method but face scalability barriers. Semantic uncertainty methods represent a genuine advance for generative models. Post-hoc calibration is unreliable under distribution shift. And the pillars are genuinely interconnected: reliable fusion requires reliable uncertainty, which benefits from meaningful disagreement signals.

Beyond their diagnostic role, uncertainty and disagreement may increasingly serve as internal control signals governing abstention, delegation, or fusion strategies in complex AI systems. Advancing AI reliability will require balancing theoretical rigour with practical constraints, ensuring that intelligent systems remain not only accurate, but aware of the limits of their own competence.

References

[1] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, “On calibration of modern neural networks,” in Proc. 34th Int. Conf. Machine Learning (ICML), PMLR, vol. 70, pp. 1321–1330, 2017.

[2] L. Kuhn, Y. Gal, and S. Farquhar, “Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation,” in Proc. Int. Conf. Learning Representations (ICLR), 2023. arXiv:2302.09664.

[3] Y. Jiang, V. Nagarajan, C. Baek, and J. Z. Kolter, “Assessing generalization of SGD via disagreement,” in Proc. Int. Conf. Learning Representations (ICLR), 2022. arXiv:2106.13799.

[4] Z. Gao, X. Jiang, X. Xu, F. Shen, Y. Li, and H. T. Shen, “Embracing unimodal aleatoric uncertainty for robust multimodal fusion,” in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), pp. 26876–26885, 2024.

[5] J. Gawlikowski et al., “A survey of uncertainty in deep neural networks,” Artificial Intelligence Review, vol. 56, pp. 1513–1589, 2023. arXiv:2107.03342.

[6] M. Abdar et al., “A review of uncertainty quantification in deep learning: Techniques, applications and challenges,” Information Fusion, vol. 76, pp. 243–297, 2021. arXiv:2011.06225.

[7] E. Hüllermeier and W. Waegeman, “Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods,” Machine Learning, vol. 110, no. 3, pp. 457–506, 2021. arXiv:1910.09457.

[8] A. Der Kiureghian and O. Ditlevsen, “Aleatoric or epistemic? Does it matter?” Structural Safety, vol. 31, no. 2, pp. 105–112, 2009.

[9] R. M. Neal, Bayesian Learning for Neural Networks, Lecture Notes in Statistics, vol. 118. New York, NY: Springer, 1996.

[10] A. Kendall and Y. Gal, “What uncertainties do we need in Bayesian deep learning for computer vision?” in Advances in Neural Information Processing Systems (NeurIPS), pp. 5574–5584, 2017. arXiv:1703.04977.

[11] T. Gneiting and A. E. Raftery, “Strictly proper scoring rules, prediction, and estimation,” J. Amer. Statistical Assoc., vol. 102, no. 477, pp. 359–378, 2007.

[12] Y. Ovadia, E. Fertig, J. Ren, et al., “Can you trust your model’s uncertainty? Evaluating predictive uncertainty under dataset shift,” in Advances in Neural

Information Processing Systems (NeurIPS), vol. 32, 2019. arXiv:1906.02530.

[13] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra, “Weight uncertainty in neural networks,” in Proc. 32nd Int. Conf. Machine Learning (ICML), PMLR, vol. 37, pp. 1613–1622, 2015. arXiv:1505.05424.

[14] Y. Gal and Z. Ghahramani, “Dropout as a Bayesian approximation: Representing model uncertainty in deep learning,” in Proc. 33rd Int. Conf. Machine Learning (ICML), PMLR, vol. 48, pp. 1050–1059, 2016. arXiv:1506.02142.

[15] A. Immer et al., “Scalable marginal likelihood estimation for model selection in deep learning,” in Proc. 38th Int. Conf.

Machine Learning (ICML), PMLR, vol. 139, 2021. arXiv:2104.04975.

[16] B. Lakshminarayanan, A. Pritzel, and C. Blundell, “Simple and scalable predictive uncertainty estimation using deep ensembles,” in Advances in Neural Information Processing Systems (NeurIPS), pp. 6402–6413, 2017. arXiv:1612.01474.

[17] Y. Wen, D. Tran, and J. Ba, “BatchEnsemble: An alternative approach to efficient ensemble and lifelong learning,” in Proc. Int. Conf. Learning Representations (ICLR), 2020.

[18] X. Wang, J. Wei, D. Schuurmans, et al., “Self-consistency improves chain of thought reasoning in language models,” in Proc. Int. Conf. Learning Representations (ICLR), 2022. arXiv:2203.11171.

[19] P. Manakul, A. Liusie, and M. J. F. Gales, “SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models,” in Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP), pp. 9004–9017, 2023. arXiv:2303.08896.

[20] S. Kadavath et al., “Language models (mostly) know what they know,” arXiv preprint arXiv:2207.05221, 2022.

[21] A. N. Angelopoulos and S. Bates, “A gentle introduction to conformal prediction and distribution-free uncertainty quantification,” Foundations and Trends in Machine Learning, vol. 16, no. 4, pp. 494–591, 2023. arXiv:2107.07511.

[22] Y. Romano, E. Patterson, and E. J. Candès, “Conformalized quantile regression,” in Advances in Neural Information Processing Systems (NeurIPS), 2019. arXiv:1905.03222.

[23] A. N. Angelopoulos, S. Bates, J. Malik, and M. I. Jordan, “Uncertainty sets for image classifiers using conformal prediction,” in Proc. Int. Conf. Learning Representations (ICLR), 2021. arXiv:2009.14193.

[24] M. Sensoy, L. Kaplan, and M. Kandemir, “Evidential deep learning to quantify classification uncertainty,” in Advances in Neural Information Processing Systems (NeurIPS), vol. 31, pp. 3179–3189, 2018. arXiv:1806.01768.

[25] A. Amini, W. Schwarting, A. Soleimany, and D. Rus, “Deep evidential regression,” in Advances in Neural Information Processing Systems (NeurIPS), vol. 33, pp. 14927–14937, 2020. arXiv:1910.02600.

[26] B. Charpentier, D. Zügner, and S. Günnemann, “Posterior network: Uncertainty estimation without OOD samples via density-based pseudo-counts,” in Advances in Neural Information Processing Systems (NeurIPS), 2020. arXiv:2006.09239.

[27] D. Ulmer, C. Hardmeier, and J. Frellsen, “Prior and posterior networks: A survey on evidential deep learning methods for uncertainty estimation,” Trans. Machine Learning Research (TMLR), 2023. arXiv:2110.03051.

[28] S. Farquhar, J. Kossen, L. Kuhn, and Y. Gal, “Detecting hallucinations in large language models using semantic entropy,” Nature, vol. 630, no. 8017, pp. 625–630, 2024.

[29] D. Hendrycks and K. Gimpel, “A baseline for detecting misclassified and out-of-distribution examples in neural networks,” in Proc. Int. Conf. Learning Representations (ICLR), 2017. arXiv:1610.02136.

[30] S. Liang, Y. Li, and R. Srikant, “Enhancing the reliability of out-of-distribution image detection in neural networks,” in Proc. Int. Conf. Learning Representations (ICLR), 2018. arXiv:1706.02690.

[31] W. Liu, X. Wang, J. D. Owens, and Y. Li, “Energy-based out-of-distribution detection,” in Advances in Neural Information Processing Systems (NeurIPS), 2020. arXiv:2010.03759.

[32] G. Shafer, A Mathematical Theory of Evidence. Princeton, NJ: Princeton Univ. Press, 1976.

[33] R. E. Kalman, “A new approach to linear filtering and prediction problems,” J. Basic Engineering, vol. 82, pp. 35–45, 1960.

[34] C. K. Chow, “On optimum error and reject tradeoff,” IEEE Trans. Information Theory, vol. 16, no. 1, pp. 41–46, 1970.

[35] Y. Geifman and R. El-Yaniv, “SelectiveNet: A deep neural network with an integrated reject option,” in Proc. 36th Int. Conf. Machine Learning (ICML), PMLR, vol. 97, pp. 2151–2159, 2019.

[36] J. Z. Liu, Z. Lin, S. Padhy, D. Tran, T. Bedrax-Weiss, and B. Lakshminarayanan, “Simple and principled uncertainty estimation with deterministic deep learning via distance awareness,” in Advances in Neural Information Processing Systems (NeurIPS), 2020. arXiv:2006.10108.

[37] C. E. Rasmussen and C. K. I. Williams, Gaussian Processes for Machine Learning. Cambridge, MA: MIT Press, 2006.

[38] C. M. Bishop, “Mixture density networks,” Technical Report, Aston University, 1994.

[39] D. J. C. MacKay, “Bayesian methods for adaptive models,” Ph.D. thesis, California Institute of Technology, 1992.

[40] S. Lin, J. Hilton, and O. Evans, “Teaching models to express their uncertainty in words,” Trans. Machine Learning Research (TMLR), 2022. arXiv:2205.14334.

[41] M. Xiong, Z. Hu, X. Lu, Y. Li, J. Fu, J. He, and B. Hooi, “Can LLMs express their uncertainty? An empirical evaluation of confidence elicitation in LLMs,” in Proc. Int. Conf. Learning Representations (ICLR), 2024. arXiv:2306.13063.

[42] B. Charpentier, O. Borchert, D. Zügner, S. Geisler, and S. Günnemann, “Natural posterior network: Deep Bayesian predictive uncertainty for exponential family distributions,” in Proc. Int. Conf. Learning Representations (ICLR), 2022. arXiv:2105.04471.

[43] A. G. Wilson, Z. Hu, R. Salakhutdinov, and E. P. Xing, “Deep kernel learning,” in Proc. 19th Int. Conf. Artificial Intelligence and Statistics (AISTATS), PMLR, vol. 51, 2016. arXiv:1511.02222.

[44] N. Felicioni, L. Maystre, S. Ghiassian, and K. Ciosek, “On the importance of uncertainty in decision-making with large language models,” arXiv preprint arXiv:2404.02649, 2024.

[45] J. Hensman, A. G. de G. Matthews, and Z. Ghahramani, “Scalable variational Gaussian process classification,” in Proc. 18th Int. Conf. Artificial Intelligence and Statistics (AISTATS), PMLR, vol. 38, pp. 351–360, 2015. arXiv:1411.2005.

[46] M. J. Page, J. E. McKenzie, P. M. Bossuyt, et al., “The PRISMA 2020 statement: An updated guideline for reporting systematic reviews,” BMJ, vol. 372, p. n71, 2021. https://doi.org/10.1136/bmj.n71

[47] J. Pineau, P. Vincent-Lamarre, K. Sinha, V. Larivière, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and H. Larochelle, “Improving reproducibility in machine learning research (a report from the NeurIPS 2019 reproducibility program),” J. Machine Learning Research, vol. 22, no. 164, pp. 1–20, 2021. arXiv:2003.12206.

[48] V. Vovk, A. Gammerman, and G. Shafer, Algorithmic Learning in a Random World. New York, NY: Springer, 2005.

[49] J. Yang, K. Zhou, Y. Li, and Z. Liu, “Generalized out-of-distribution detection: A survey,” Int. J. Computer Vision, vol. 132, 2024. arXiv:2110.11334.

[50] Q. Zhang, H. Wu, C. Zhang, Q. Hu, H. Fu, J. T. Zhou, and X. Peng, “Provable dynamic fusion for low-quality multimodal data,” in Proc. 40th Int. Conf. Machine Learning (ICML), PMLR, vol. 202, pp. 41753–41769, 2023.

[51] E. Daxberger, A. Kristiadi, A. Immer, R. Eschenhagen, M. Bauer, and P. Hennig, “Laplace redux — Effortless Bayesian deep learning,” in Advances in Neural Information Processing Systems (NeurIPS), vol. 34, 2021. arXiv:2106.14806.

[52] A. Kirsch and Y. Gal, “A note on ‘Assessing generalization of SGD via disagreement’,” Trans. Machine Learning Research (TMLR), 2022.

[53] T. Abe, E. K. Buchanan, G. Pleiss, and J. P. Cunningham, “Pathologies of predictive diversity in deep ensembles,” Trans. Machine Learning Research (TMLR), 2024. arXiv:2302.00704.

[54] J. C. Peterson, R. M. Battleday, T. L. Griffiths, and O. Russakovsky, “Human uncertainty makes classification more robust,” in Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV), 2019.

[55] A. N. Uma, T. Fornaciari, D. Hovy, S. Paun, B. Plank, and M. Poesio, “Learning from disagreement: A survey,” J. Artificial Intelligence Research, vol. 72, pp. 1385–1470, 2021.

[56] B. Plank, “The ‘problem’ of human label variation: On ground truth in data, modeling and evaluation,” in Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP), 2022.

[57] Y. Du, S. Li, A. Torralba, J. B. Tenenbaum, and I. Mordatch, “Improving factuality and reasoning in language models through multiagent debate,” arXiv preprint arXiv:2305.14325, 2023.

[58] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, and J. Dean, “Outrageously large neural networks: The sparsely-gated mixture-of-experts layer,” in Proc. Int. Conf. Learning Representations (ICLR), 2017.