Confidence as Control: A Survey of Confidence Utilization in Large Language Models

Anonymous ACL submission

Abstract

Confidence in LLMs is often studied through uncertainty estimation and calibration. We survey a complementary perspective: confidence as a control signal that governs system behavior. We organize **confidence utilization** across the LLM lifecycle: (i) training (data selection, loss weighting, self-training, and preference optimization); (ii) inference (candidate selection, adaptive computation, and confidence-guided contrastive decoding); and (iii) deployment (cost-aware routing and cascading, RAG control (retrieval triggering, context filtering, and parametric–retrieval arbitration), and risk-aware abstention and monitoring. We unify these techniques into a framework that turns confidence into system decisions, with implications for efficiency and reliability.

1 Introduction

Confidence in Large Language Models (LLMs) bridges the gap between internal uncertainty and actionable system behavior. A model that knows what it doesn't know is useful; a system that acts on this knowledge—retrieving when uncertain, deferring when unreliable, focusing learning where needed—is transformative. This survey concerns the latter. Recent work has demonstrated that LLMs can express calibrated uncertainty through verbalized probabilities (Tian et al., 2023; Kadavath et al., 2022), that semantic entropy over meaning clusters detects hallucinations (Farquhar et al., 2024), and that self-consistency among sampled responses provides robust confidence signals (Wang et al., 2022; Manakul et al., 2023). These estimation advances have successfully equipped models with the ability to signal their own reliability.

At the same time, several surveys have examined confidence and uncertainty in LLMs, predominantly focusing on estimation and calibration. Geng et al. (2024) provide a foundational taxonomy of confidence estimation and calibration techniques, while Shorinwa et al. (2025) and Liu et al. (2025b) offer comprehensive coverage of uncertainty quantification methods and their applications. Xiong et al. (2023) and Xie et al. (2024) address the specific challenges of model-agnostic settings where model internals are inaccessible. These surveys answer: how can we obtain high-quality confidence scores? Yet a good confidence score is not the end goal, it is the prerequisite. A systematic treatment of how confidence should actually govern system behavior remains absent.

Our survey addresses the critical next step: **how should systems utilize confidence?** As illustrated in Figure 1, we propose a taxonomy[1] where confidence functions as *control*—not a passive measure of uncertainty, but an active governor that determines what to learn, how to reason, and when to defer. We trace this utilization across the full LLM lifecycle including: (i) training-time applications in data curation and alignment (§2), (ii) inference-time control over reasoning and decoding (§3), and (iii) deployment-time decisions including model selection (§4), retrieval-augmented generation (RAG) (§5), and risk management (§6). By synthesizing these distinct domains, we aim to offer a unified perspective for designing systems that act on uncertainty rather than merely quantify it.

2 Confidence-Aware Training

Confidence signals shape the training at multiple stages: selecting which data to train on, modulating how models learn from that data, and guiding preference optimization for alignment. We organize methods by where confidence intervenes: data curation (§2.1), supervised fine-tuning (§2.2), and reinforcement learning from human feedback (§2.3).

2.1 Confidence-aware data selection

Confidence-aware data selection uses model uncertainty as a proxy for data value, varying in con-

[1] Across this survey, we employ standard field notations summarized in Appendix A without explicit re-definition.

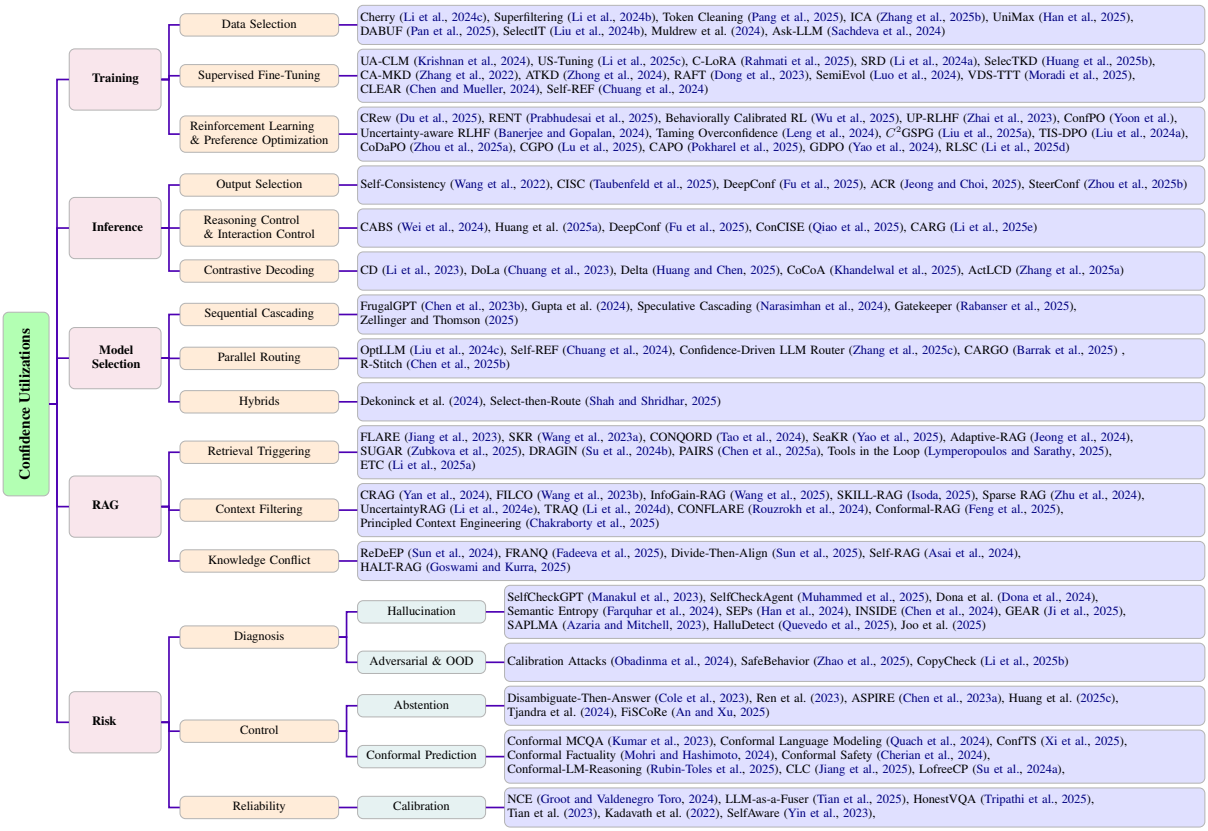

Figure 1: Taxonomy of Confidence Utilization in LLMs

fidence source (self-evaluation, external models, LLM-as-judge), granularity (sample vs. token), and selection goal (quality filtering, influence estimation, active acquisition).

Difficulty as inverse confidence. Low confidence often signals learning potential. Li et al. (2024c) define Instruction-Following Difficulty $\mathrm{IFD}(x, y) = \mathcal{L}_\theta(y|x)/\mathcal{L}_\theta(y)$, isolating instruction-following difficulty from baseline response complexity. Li et al. (2024b) show IFD rankings transfer across scales—GPT-2's orderings correlate with LLaMA2-7B/13B, enabling efficient filtering. Pang et al. (2025) extend this to token granularity by comparing base-model and clean-reference losses: tokens are retained when the base model struggles but the reference succeeds (learnable content), and filtered when both show high loss (noise).

Influence and acquisition. Zhang et al. (2025b) estimate sample utility via In-Context Approximation—conditioning on a holdout set simulates gradient steps without retraining. Han et al. (2025) frame selection as graph influence maximization with epistemic uncertainty balancing influence against redundancy; Pan et al. (2025) apply attribution in reverse, tracing harmful outputs to training samples for removal. For label-efficient settings, Muldrew et al. (2024) combine predictive entropy with preference model certainty to guide annotation acquisition; Sachdeva et al. (2024) query an LLM directly, using $\mathrm{P(YES)}$ as quality score. Liu et al. (2024b) exploit multi-granularity uncertainty (token, sentence, and model-level), selecting samples where self-assessment is confident and consistent.

2.2 Confidence-aware fine-tuning

Data selection determines what to train on; confidence also reshapes how models learn. Traditional supervised fine-tuning (SFT) treats tokens equally, yet highly predictable tokens (articles, connectives) and already-learned patterns provide diminishing learning signal. Confidence-modulated methods address this by scaling each token's contribution to the loss based on certainty. The central design question: should confident predictions receive more weight (reinforcing what the model knows) or less (focusing learning on uncertain regions)?

Confidence-weighted training. Krishnan et al. (2024) decompose uncertainty into aleatoric and epistemic components, using the epistemic term to emphasize tokens where model knowledge is lacking. Li et al. (2025c) target a different goal: training LLMs to recognize knowledge boundaries and abstain rather than hallucinate. Rahmati et al. (2025) take a Bayesian approach, conditioning LoRA adapter distributions on each input so that

uncertainty estimates become sample-specific, improving calibration in few-shot settings.

Knowledge distillation with confidence. Naive distillation treats all teacher outputs as equally reliable, but teachers are often uncertain on ambiguous tokens. Huang et al. (2025b) filter via propose-and-verify: students propose tokens, teachers accept or reject based on distribution agreement, and only accepted tokens receive distillation signal. Li et al. (2024a) curate data through student self-reflection, ranking samples by difficulty and selecting those the student can realistically learn from. Zhang et al. (2022) weight multi-teacher ensembles by per-sample confidence. Zhong et al. (2024) distinguish easy from hard tokens: easy tokens skip imitation-focused teaching while hard tokens receive richer, more diverse supervision.

Confidence in self-training. Self-training relies on model-generated pseudo-labels, making confidence essential for filtering unreliable outputs. Dong et al. (2023) generate multiple responses per prompt, rank by reward, and train only on top-scoring samples. Chen and Mueller (2024) automate curation more broadly: confidence estimates identify low-quality training pairs, which are either filtered or replaced with higher-confidence LLM-generated alternatives. Luo et al. (2024) extend rejection sampling to semi-supervised settings; Moradi et al. (2025) add verification to detect when self-generated data helps versus harms. Beyond filtering, Chuang et al. (2024) train models to output calibrated confidence tokens, enabling routing to stronger models when local confidence is low.

2.3 Confidence-aware preference optimization

Preference optimization aligns LLMs with human values, but reward models can inherit annotation noise and policies exploit blind spots. Given the objective $\max_\theta \mathbb{E}[r(x, y)] - \beta \cdot \mathrm{KL}(\pi_\theta \| \pi_{\text{ref}})$, confidence integrates as the reward itself, as an uncertainty penalty, or as granular weights over tokens, trajectories, or preference pairs.

Confidence as reward. Du et al. (2025) use token-level log-probability directly as reward, while Prabhudesai et al. (2025) apply entropy minimization for reasoning tasks. Naive confidence maximization risks reward hacking; Leng et al. (2024) counter this by calibrating reward models during PPO via confidence-augmented preference data and running-average reward thresholds.

Uncertainty penalties. Zhai et al. (2023) and Banerjee and Gopalan (2024) penalize by reward model uncertainty ($r_{\text{eff}} = r - \alpha\sigma_r$, where σ_r is reward uncertainty and α penalty strength), a pessimistic approach preferring conservative estimates. Liu et al. (2025a) integrate calibration directly into policy gradients via cross-entropy regularization aligning sequence probability with reward.

Granular weighting. At the token level, Yoon et al. focus optimization on high-uncertainty decision points while Liu et al. (2024a) weight by probability differences between contrastive models. At the trajectory level, Zhou et al. (2025a) rescale advantages by both confidence and difficulty; Lu et al. (2025) construct preference pairs from confidence-ranked outputs. At the preference level, Pokharel et al. (2025) dynamically scale DPO loss (Rafailov et al., 2023) based on relative reward margins, down-weighting ambiguous comparisons.

Self-feedback. Li et al. (2025d) use self-assessed confidence as intrinsic reward, generating pseudo-preferences without external labels. Wu et al. (2025) train calibrated abstention via Brier-score rewards (Gneiting and Raftery, 2007) that incentivize refusal below user-specified risk thresholds.

By transforming confidence from a post-hoc metric into an intrinsic training signal, these methods shift the paradigm from blind data absorption to discerning learning. This integration fosters models that do not merely mimic patterns, but actively navigate the boundaries of their own knowledge.

3 Confidence-Driven Inference

We examine how confidence is used during inference to guide LLM behavior, independent of how confidence is estimated. Existing approaches apply confidence to select among candidate outputs (§3.1), control reasoning and interaction as inference unfolds (§3.2), or guide decoding implicitly through predictive discrepancies (§3.3).

3.1 Confidence-driven output selection

When LLMs are sampled multiple times at inference, they typically produce a set $\mathcal{C}$ including candidate outputs $\{c_1, \cdots, c_n\}$ (e.g., responses or reasoning traces) rather than a single deterministic answer. Each candidate c_j is associated with a confidence score $s_j \in \mathbb{R}$, estimated by some mechanism. Inference-time output selection maps $\mathcal{C}$ to a final output $\hat{c}$. While all methods in this subsection perform output selection, they differ in how confidence is used as an inference operator, either by aggregating multiple candidates jointly or by scoring candidates and selecting one.

In confidence-weighted aggregation, the final output is determined jointly by multiple candidates,

$\hat{c} = \arg\max_{c \in \mathcal{C}} \sum_{j=1}^{n} s_j \mathbb{I}(c_j = c)$, where each candidate's contribution is weighted by its confidence score. Unlike classical self-consistency, which aggregates reasoning traces via uniform majority voting (Wang et al., 2022), several works weight candidates according to confidence, allowing infrequent but high-confidence correct answers to prevail the final decision (Taubenfeld et al., 2025; Fu et al., 2025). Confidence weighting suppresses unreliable traces and elevates correct minorities.

Unlike aggregation, no information is merged across candidates in confidence-based scoring, where inference is governed by selection rather than combination. More specifically, candidates are evaluated independently and the final output is selected as $\hat{c} = \arg\max_{c_j \in \mathcal{C}} s_j$, with confidence serving as a reliability score rather than a weight. Methods in Jeong and Choi (2025) and Zhou et al. (2025b) apply this paradigm by re-scoring multiple candidates and selecting the most confident one.

3.2 Confidence-driven control of decoding

While confidence can be applied after inference to select among completed outputs, it can also act during inference to control how reasoning and interaction unfold, where confidence functions as a control signal over the reasoning process rather than only the final decision. Consider an inference trajectory represented as a sequence of states and associated confidence scores $(z_1, s_1) \rightarrow (z_2, s_2) \rightarrow \cdots \rightarrow (z_T, s_T)$. Confidence-driven reasoning control modifies the transition policy into $z_{t+1} \sim \pi(z_{t+1}|z_{1;t}, s_{1:t})$, allowing confidence to govern continuation, refinement, and belief maintenance across steps or turns.

Adaptive computation and early stopping. Some methods use confidence to allocate inference-time computation by continuing reasoning only when confidence remains below a threshold. Huang et al. (2025a) terminate sampling once confidence is sufficiently high, reducing inference cost without sacrificing accuracy, while Fu et al. (2025) allocate deeper reasoning only when confidence remains low. Qiao et al. (2025) apply this principle at the chain-of-thought level, terminating or compressing reasoning once confidence exceeds a threshold and stabilizes ($|s_t - s_{t-1}| \leq \epsilon$). Across these approaches, confidence serves as a controller that prevents unnecessary computation.

Refinement. Confidence can guide where refinement occurs by identifying unreliable intermediate states ($s_t \leq \tau$) and selectively refining only these parts. Wei et al. (2024) selectively regenerate low-confidence substructures in structured generation, while Fu et al. (2025) prune low-confidence reasoning paths and re-explores alternatives. In both cases, confidence directs refinement efforts.

Multi-turn stability and interaction control. In multi-turn interactions, confidence can also regulate behavior across turns by discouraging unnecessary revision of high-confidence earlier responses. Li et al. (2025e) incorporate confidence into the dialogue state by extracting specific tokens to estimate response certainty, conditioning future responses on both prior text and associated confidence to maintain consistency across interactions.

3.3 Confidence-driven contrastive decoding

Contrastive decoding methods exploit disagreement between predictive views as an implicit confidence signal during inference, rather than estimating confidence explicitly. Depending on how predictive views are constructed (across contexts, layers, or models), different forms of disagreement arise, but in all cases they reflect the relative stability of token support. At the decoding step t (next-token prediction given the current prefix), two views produce two sets of logits $\text{logit}_t^{(a)}$ and $\text{logit}_t^{(b)}$, where view a is expected to be more reliable by construction. A contrastive logit is defined as $\tilde{\text{logit}}_t = \text{logit}_t^{(a)} - \lambda \text{logit}_t^{(b)}$, where $\lambda > 0$ controls the strength of penalizing the less reliable view. Decoding then samples the next token from the distribution $\pi_t \propto \exp(\tilde{\text{logit}}_t)$, favoring tokens whose support is stable under the more reliable view and weak under the less reliable one.

To build these predictive views, context–parametric methods first contrast context-conditioned predictions with the model's parametric prior: Huang and Chen (2025) down-weight tokens that remain likely under random context masking, while Khandelwal et al. (2025) adaptively balances parametric and contextual distributions based on their conflict, treating reliance on informative context as confidence. Alternatively, intra-model methods derive contrast from different transformer layers: Chuang et al. (2023) favors tokens whose support strengthens from shallow to deep layers, and Zhang et al. (2025a) learns when such contrast should be applied, encoding confidence as consistency across representational depth. Finally, the inter-model method proposed by Li et al. (2023) contrasts predictions from an expert and a weaker model, amplifying tokens preferred by the expert and using model-level

disagreement for token reliability.

At inference time, confidence can function as a signal that governs output selection, reasoning dynamics, and decoding behavior, enabling more reliable generation without retraining.

4 Confidence-Guided Model Selection

Model selection addresses a practical tension in LLM deployment: stronger models tend to be more accurate but also more costly. Confidence provides a principled control signal for model selection, enabling systems to decide when a cheaper model is sufficient and when to invoke a stronger model. In this section, we review confidence-guided model-selection architectures (§4.1) and the control policies (§4.2) that operationalize confidence for routing, cascading, and their hybrids.

4.1 Confidence-guided selection architectures

Prior work instantiates confidence-guided model selection via (i) sequential cascading, (ii) parallel routing, or (iii) hybrids that combine both.

Cascading. Sequential cascading invoke models in increasing strength and use confidence to decide whether to stop or escalate. This paradigm is analyzed in Gupta et al. (2024), operationalized for cost control in Chen et al. (2023b), accelerated via speculative-decoding-based cascading in Narasimhan et al. (2024), improved by confidence-aware tuning of smaller models in Rabanser et al. (2025), and optimized through probabilistic threshold tuning in Zellinger and Thomson (2025).

Routing. Parallel routing selects one model (or a small subset) without sequential escalation, using confidence to choose among candidates and to manage routing overhead. Examples include routing as query-to-model assignment (Liu et al., 2024c), confidence-token interfaces for downstream decisions (Chuang et al., 2024), uncertainty-driven routing policies (Zhang et al., 2025c), and confidence-aware prediction frameworks for scalable selection (Barrak et al., 2025).

Hybrids. Hybrid systems combine routing and cascading by using confidence to narrow candidates and then escalating selectively. Representative methods include cascade routing as a unified policy class (Dekoninck et al., 2024) and select-then-route pipelines that shortlist models before adaptive escalation (Shah and Shridhar, 2025).

4.2 Confidence-guided control policies

Complementary to architectural instantiations, we summarize policy primitives that map a confidence/quality signal to actions (escalate, select a model, or allocate additional compute) and can be embedded in any architecture.

Thresholding / abstain–escalate. The canonical policy is confidence-gated deferral: a cheaper model answers when confidence is high and defers otherwise. Gupta et al. (2024) show that naive sequence-level confidence can be misleading and propose token-level aggregation (e.g., quantiles) and learned decision functions for improved cost–quality trade-offs. Rabanser et al. (2025) strengthen deferral by training smaller models to be confident when correct and to defer when uncertain. Zellinger and Thomson (2025) address threshold tuning directly by modeling joint calibrated confidences across cascade stages and optimizing thresholds via differentiable error-cost objectives.

Selection among candidates. A second family uses confidence to select among candidate models. Chuang et al. (2024) introduce confidence tokens to support routing and rejection, Zhang et al. (2025c) design uncertainty-driven routers, and Barrak et al. (2025) scale to large candidate sets by predicting likely winners from model-judged comparisons and using predictor uncertainty to control routing overhead.

Allocate compute adaptively. Finally, some methods allocate compute adaptively during inference: Chen et al. (2025b) switch from a small to a large model when confidence drops during reasoning; Narasimhan et al. (2024) combine cascading with speculative decoding via confidence-driven acceptance or deferral to reduce decoding-time compute.

In summary, confidence-guided routing and cascading implement policies that map queries (and sometimes intermediate generation states) to model-invocation decisions, thereby navigating cost–quality trade-offs (Chen et al., 2023b; Dekoninck et al., 2024). Recent token- or step-level extensions further suggest that confidence can govern not only which model to use, but also when to hand off computation during generation (Narasimhan et al., 2024; Chen et al., 2025b).

5 Confidence-Gated RAG Systems

We focus on the use of confidence: given a reliable signal, what mechanisms translate confidence into actionable RAG behavior? We organize existing work into three categories: adaptive retrieval triggering (§5.1), context filtering and selection (§5.2), and knowledge conflict resolution (§5.3).

5.1 Adaptive retrieval triggering

RAG faces an efficiency-accuracy trade-off: unconditional retrieval wastes computation, while rigid refusal limits models to stale parametric knowledge. Confidence resolves this by transforming static retrieval policies into dynamic resource allocation.

Whether to retrieve. Most methods implement confidence-gated retrieval. Jiang et al. (2023) generate a draft sentence and trigger retrieval when it contains low-probability tokens, using those spans as queries. Wang et al. (2023a) elicit self-knowledge to decide retrieval necessity, while Tao et al. (2024) align verbalized confidence with response quality so that confidence reliably gates retrieval. Beyond binary gating, Jeong et al. (2024) and Zubkova et al. (2025) route queries across no retrieval, single-step, or multi-hop retrieval based on predicted complexity or entropy, and Yao et al. (2025) extract self-aware uncertainty from activations to select retrieval strategies.

What to retrieve. Confidence also guides what to retrieve: Su et al. (2024b) detect increasing uncertainty by combining token-level entropy, attention-derived influence, and semantic significance, forming queries from context-wide self-attention rather than only recent tokens. Besides, some methods treat retrieval as expected utility: Chen et al. (2025a) compare the parametric answer to a self-generated contextual probe—convergence bypasses retrieval, divergence triggers it; Asai et al. (2024) learn special tokens whose probabilities encode retrieval necessity and critique judgments; Lymperopoulos and Sarathy (2025) model joint uncertainty over the LLM and external tools to support retrieve-or-abstain decisions.

When to retrieve. Pre-generation methods (Wang et al., 2023a; Jeong et al., 2024; Tao et al., 2024) assess confidence from the query or a probe before answering. Mid-generation methods (Jiang et al., 2023; Su et al., 2024b) interleave retrieval when uncertainty spikes during decoding. Yan et al. (2024) operate post-retrieval: a lightweight evaluator scores retrieved evidence confidence and triggers corrective actions—including web search escalation—when retrieval quality is insufficient. Li et al. (2025a) further show that retrieval timing can be determined by exploiting trends (first- and second-order entropy differences) to retrieve preemptively, before errors propagate.

5.2 Confidence-based context filtering

After retrieval is triggered, the challenge becomes context allocation: selecting which evidence to include within limited context budgets. Confidence serves as a utility signal, estimating whether retrieved content would improve generation rather than merely match the query.

Utility-scored filtering. A common pattern is to compute (or learn) a scalar utility score for each evidence unit and retain only high-utility content. Wang et al. (2023b) construct oracle usefulness signals from lexical overlap and conditional cross-mutual information, the confidence increase when a span is provided, then train a sentence-level filter that removes distracting content at test time. Yan et al. (2024) use a lightweight evaluator that scores retrieved-set quality as CORRECT, INCORRECT, or AMBIGUOUS, triggering corrective actions (keep, refine, or escalate to web search) and decomposing documents to focus on key information.

Information-gain filtering. Absolute relevance can be misleading: an on-topic document may add noise rather than signal. Wang et al. (2025) define Document Information Gain as the change in generation confidence with vs. without a document, training a reranker to prioritize evidence that meaningfully shifts confidence. Isoda (2025) filter at sentence level using a self-knowledge confidence score: the model outputs a scalar "know/unknown" probability for each candidate sentence, and an RL policy learns to keep sentences that increase this self-knowledge (and drop the rest). Zhu et al. (2024) assign each document a relevance confidence from control-token judgments, retain only high-score docs, and decode using only their KV caches to reduce noise and latency. Li et al. (2024e) use SNR-based span uncertainty to estimate chunk similarity in long-context retrieval, improving calibration and robustness to chunking artifacts and distribution shift.

Conformal filtering. Conformal prediction calibrates raw confidence scores—retrieval similarities or generation probabilities—into threshold rules with coverage guarantees. Li et al. (2024d) calibrate retrieval scores to form passage sets and generation confidence to form answer sets, composing both for end-to-end guarantees. Rouzrokh et al. (2024) calibrate similarity scores to select context at a user-specified error rate. Chakraborty et al. (2025) show that calibrating any scoring function yields a principled filter that reduces retained context by 2-3 times while preserving coverage. Feng et al. (2025) apply conformal calibration post-hoc to sub-claim factuality scores.

5.3 Confidence for knowledge conflict

Retrieved context may contradict parametric knowledge; confidence quantifies the reliability of each source to arbitrate the conflict.

Confidence signals for conflict. Recent work turns "conflict" into measurable confidence signals rather than answer disagreement. Sun et al. (2024) detect RAG hallucinations via mechanistic interpretability, computing separate scores for parametric-knowledge utilization and external-context integration; hallucinations correlate with FFN dominance coupled with copying-head failure. Fadeeva et al. (2025) separate faithfulness (entailed by retrieved context) from factuality (true in the world) and route uncertainty quantification accordingly, so "unsupported by retrieval" is not conflated with "false." Asai et al. (2024) formalize support at inference time: learned reflection-token probabilities (relevance, support, usefulness) yield calibrated, inspectable confidence over whether each continuation is grounded in retrieved evidence.

Resolution policies. Once confidence exposes a conflict, it must drive an action. Sun et al. (2025) use confidence to partition queries by whether they fall within the model's parametric knowledge boundary and the retrieval boundary; DPO then trains quadrant-specific behaviors, answering when at least one source suffices and abstaining when both fail. Goswami and Kurra (2025) apply calibrated NLI-ensemble scores as a post-hoc arbitration layer, abstaining when entailment confidence falls below a threshold. Soudani et al. (2025) diagnose why standard uncertainty estimators fail in RAG and propose an axiomatic calibration that conditions uncertainty jointly on model and evidence.

Confidence turns RAG into a closed-loop policy: it gates retrieval, allocates context by marginal utility, and arbitrates conflicts by triggering resolution, escalation, or abstention strategies.

6 Confidence-Based Risk Management

While previous sections utilize confidence to enhance performance and efficiency, this section focuses on risk control and reliability. We approach risk management through a confidence-based three-stage framework: first diagnosing anomalies such as hallucinations and adversarial attacks (§6.1), then controlling risk through mechanisms like abstention and conformal prediction (§6.2), and finally ensuring the reliability of the confidence signal itself to reduce overconfidence (§6.3).

6.1 Confidence-based risk diagnosis

We categorize diagnostic mechanisms by the source of risk: detecting internal model hallucinations and flagging external input anomalies.

Hallucination detection. Hallucinations represent a misalignment between the model's generated output and factual reality (Rawte et al., 2023). Confidence scores can be applied to this task as diagnostic signals. A primary application involves checking output consistency: works like Manakul et al. (2023) and Muhammed et al. (2025) use self-consistency scores to measure disagreement across independently sampled responses as a proxy for hallucination. This strategy is further adapted to filter errors in domain-specific applications like autonomous driving (Dona et al., 2024). To address the ambiguity of phrasing, Farquhar et al. (2024) introduce Semantic Entropy over sampled meaning clusters, while Han et al. (2024) efficiently approximate it via linear probes to avoid sampling cost. Beyond sampling, detection mechanisms exploit internal state anomalies, utilizing the EigenScore (Chen et al., 2024) to measure the semantic consistency in the embedding space, or training classifiers on hidden layer activations (Azaria and Mitchell, 2023) and various confidence-derived features (Quevedo et al., 2025). Other approaches employ active diagnosis of hallucination by measuring confidence fluctuations under perturbed queries (Joo et al., 2025) or intervening on learned verbal uncertainty features (Ji et al., 2025).

Adversarial and OOD detection. Confidence signals can serve as indicators for adversarial and out-of-distribution (OOD) anomalies. For instance, (Obadinma et al., 2024) utilizes confidence-aware training to mitigate calibration attacks, while SafeBehavior (Zhao et al., 2025) computes a "safety confidence score" to actively trigger self-revision processes during jailbreak attempts. In the context of OOD detection, confidence patterns are utilized to trace data provenance; for example (Li et al., 2025b) aggregates confidence scores to detect membership inference risks by recognizing previously seen training data.

6.2 Confidence-based risk control

Moving from detection to intervention, we review how confidence thresholds are employed to gate unreliable generations and bound output error rates.

Abstention. Abstention refers to the refusal of an LLM to provide an answer (Wen et al., 2025). A common confidence-driven approach uses a gatekeeping mechanism to abstain from answering

when confidence is below a threshold τ. In post-hoc thresholding, confidence scores are applied directly to frozen models to judge response quality. A good confidence score for abstention should focus on tracking correctness rather than fluency or typicality. Following this principle, Cole et al. (2023) applies self-consistency for abstention and and shows it yields better performance than log-likelihood, while Ren et al. (2023) utilizes LLM self-evaluation scores to better discriminate low-quality generations. Alternatively, Huang et al. (2025c) probes internal activations to extract a scalar uncertainty feature specifically optimized for this thresholding decision.

Moving from inference-time filtering to model adaption, recent works focus on teaching the model when to abstain, applying confidence scores as supervisory signals. Chen et al. (2023a) fine-tunes LLMs with labeled answers to obtain a confidence score that combines self-evaluation and normalized likelihood for selective prediction. To further reduce label reliance, Tjandra et al. (2024) employs semantic entropy as a pseudo-label to enable label-free abstention fine-tuning, while An and Xu (2025) uses semantic confidence as a reward signal via RL to optimize the model's boundary recognition.

Conformal prediction. Conformal prediction (CP) (Shafer and Vovk, 2008) addresses the limitation of heuristic thresholds by transforming confidence scores into statistically valid prediction sets, ensuring that on expectation a correct[2] answer is included with a user-specified probability, or more broadly, that a specific risk measure is controlled (Angelopoulos et al., 2024). The application of CP in LLMs can be distinguished by the specific confidence signals used to construct these sets. Early approaches relied on logit-based scores: Kumar et al. (2023) applies CP to multiple choice questions using logit scores of the choices, while Quach et al. (2024) utilizes normalized likelihoods to tune sampling parameters for open-ended generation.

However, recognizing that raw logits often fails to capture semantic correctness, recent methods have shifted toward self-consistency and semantic scores. Mohri and Hashimoto (2024) argues that sets of answers are less useful than factual sub-claims, employing self-consistency and LLM-as-a-judge scores to guarantee all selected sub-claims are factual with user-specified probability. This approach is further refined by Cherian et al. (2024), which exploits the heterogeneity in score quality across prompts to adaptively adjust thresholds. To ensure coherence in reasoning, Rubin-Toles et al. (2025) extends these scores to model logical dependencies via deductibility graphs. For black-box settings, Su et al. (2024a) demonstrates that API-accessible scores like semantic similarity can serve as effective non-conformity scores. Finally, to improve the efficiency of these sets, Xi et al. (2025) demonstrates that optimizing confidence distribution via temperature scaling yields tighter coverage.

[2]In LLMs, correctness is task-defined: matching ground-truth labels for classification, or using proxies for open-ended outputs (e.g. semantic match or verifier pass).

6.3 Ensuring the reliability of confidence

The efficacy of risk frameworks depends on the reliability of confidence signals, yet LLMs frequently exhibit significant overconfidence and miscalibration (Groot and Valdenegro Toro, 2024; Tian et al., 2025). To resolve this, elicitation strategies are applied to extract more calibrated scores: Kadavath et al. (2022) established that LLMs can accurately estimate their own correctness via the logit probability $P(\text{True})$ when specifically prompted, while Tian et al. (2023) further showed that soliciting such explicit probabilities or top-k options yields significantly better calibration than open-ended verbalization. For existing scores, post-hoc calibration utilizes specialized confidence metrics like the TH-score (Tian et al., 2025) and Net Calibration Error (Groot and Valdenegro Toro, 2024) to explicitly penalize the gap between confidence and accuracy.

In summary, confidence-based risk management transforms the signal from a passive metric into a robust guardrail that detects errors, bounds risks, and ensures system reliability.

7 Conclusion

This survey presents a unified view of confidence as a control signal that governs LLM behavior across the full system lifecycle. During training, it guides data selection, loss weighting, and preference optimization; during inference, it supports output selection, adaptive computation, and contrastive decoding; and in deployment, it enables cost-aware routing, confidence-gated RAG, and risk control via abstention and conformal prediction. Across these settings, confidence consistently turns discrete heuristics—what to learn, when to retrieve, whether to answer—into more principled allocation of computation and evidence. Reliable confidence is therefore not an end, but a prerequisite for systems that act on uncertainty by deferring, retrieving, or abstaining when warranted.

Limitations

This survey focuses on **confidence utilization** as a control signal rather than confidence estimation or calibration. Readers seeking comprehensive coverage of uncertainty quantification methods should consult complementary surveys.

We primarily survey English-language literature and methods evaluated on English benchmarks; confidence utilization in multilingual and low-resource settings remains comparatively underexplored.

Control policies inherit failure modes of the underlying confidence signal (miscalibration, bias, brittleness under distribution shift or adversarial prompting), which can lead to unsafe acceptance, premature deferral, or unnecessary escalation.

Many methods add latency or compute (sampling, verifiers, multi-stage retrieval), and formal guarantees (e.g., conformal prediction) depend on assumptions and proxies for correctness that may not hold under deployment drift.

Implication. Taken together, these limitations suggest that confidence-as-control should be viewed as a systems design paradigm rather than a drop-in solution: robust deployment requires validating confidence reliability under realistic shift, aligning proxy correctness with application risk, and accounting for compute and human factors alongside model accuracy.

Acknowledgment on AI Assistance.

GPT 5.2, Claude Opus 4.5, and Gemini 3 Pro were used solely to polish the language of this manuscript (e.g., improving grammar, clarity, and style). These tools did not generate or alter the scientific content, interpretations, results, or conclusions. All content was reviewed, verified, and approved by the authors, who take full responsibility for the accuracy and integrity of the work.

References

Hao An and Yang Xu. 2025. Teaching llms to abstain via fine-grained semantic confidence reward. *arXiv preprint arXiv:2510.24020*.

Anastasios Nikolas Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. 2024. Conformal risk control. In *The Twelfth International Conference on Learning Representations*.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-rag: Learning to retrieve, generate, and critique through self-reflection.

Amos Azaria and Tom Mitchell. 2023. The internal state of an llm knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 967–976.

Debangshu Banerjee and Aditya Gopalan. 2024. Towards reliable alignment: Uncertainty-aware rlhf. *arXiv preprint arXiv:2410.23726*.

Amine Barrak, Yosr Fourati, Michael Olchawa, Emna Ksontini, and Khalil Zoghlami. 2025. Cargo: A framework for confidence-aware routing of large language models. *arXiv preprint arXiv:2509.14899*.

Debashish Chakraborty, Eugene Yang, Daniel Khashabi, Dawn Lawrie, and Kevin Duh. 2025. Principled context engineering for rag: Statistical guarantees via conformal prediction. *arXiv preprint arXiv:2511.17908*.

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

Jiefeng Chen, Jinsung Yoon, Sayna Ebrahimi, Sercan Arik, Tomas Pfister, and Somesh Jha. 2023a. Adaptation with self-evaluation to improve selective prediction in llms. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5190–5213.

Jiuhai Chen and Jonas Mueller. 2024. Automated data curation for robust language model fine-tuning. *arXiv preprint arXiv:2403.12776*.

Lingjiao Chen, Matei Zaharia, and James Zou. 2023b. Frugalgpt: How to use large language models while reducing cost and improving performance. *arXiv preprint arXiv:2305.05176*.

Wang Chen, Guanqiang Qi, Weikang Li, Yang Li, Deguo Xia, and Jizhou Huang. 2025a. Pairs: Parametric-verified adaptive information retrieval and selection for efficient rag. *arXiv preprint arXiv:2508.04057*.

Zhuokun Chen, Zeren Chen, Jiahao He, Lu Sheng, Mingkui Tan, Jianfei Cai, and Bohan Zhuang. 2025b. R-stitch: Dynamic trajectory stitching for efficient reasoning. *arXiv preprint arXiv:2507.17307*.

John Cherian, Isaac Gibbs, and Emmanuel Candes. 2024. Large language model validity via enhanced conformal prediction methods. *Advances in Neural Information Processing Systems*, 37:114812–114842.

Yu-Neng Chuang, Prathusha Kameswara Sarma, Parikshit Gopalan, John Boccio, Sara Bolouki, Xia Hu, and Helen Zhou. 2024. Learning to route llms with confidence tokens. *arXiv preprint arXiv:2410.13284*.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. 2023. Dola: Decoding by contrasting layers improves factuality in large language models. *arXiv preprint arXiv:2309.03883*.

Jeremy Cole, Michael Zhang, Dan Gillick, Julian Eisenschlos, Bhuwan Dhingra, and Jacob Eisenstein. 2023. Selectively answering ambiguous questions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 530–543.

Jasper Dekoninck, Maximilian Baader, and Martin Vechev. 2024. A unified approach to routing and cascading for llms. *arXiv preprint arXiv:2410.10347*.

Malsha Ashani Mahawatta Dona, Beatriz Cabrero-Daniel, Yinan Yu, and Christian Berger. 2024. Llms can check their own results to mitigate hallucinations in traffic understanding tasks. In *IFIP International Conference on Testing Software and Systems*, pages 114–130. Springer.

Hanze Dong, Wei Xiong, Deepanshu Goyal, Yihan Zhang, Winnie Chow, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, and Tong Zhang. 2023. Raft: Reward ranked finetuning for generative foundation model alignment. *arXiv preprint arXiv:2304.06767*.

He Du, Bowen Li, Chengxing Xie, Chang Gao, Kai Chen, and Dacheng Tao. 2025. Confidence as a reward: Transforming llms into reward models. *arXiv preprint arXiv:2510.13501*.

Ekaterina Fadeeva, Aleksandr Rubashevskii, Roman Vashurin, Shehzaad Dhuliawala, Artem Shelmanov, Timothy Baldwin, Preslav Nakov, Mrinmaya Sachan, and Maxim Panov. 2025. Faithfulness-aware uncertainty quantification for fact-checking the output of retrieval augmented generation. *arXiv preprint arXiv:2505.21072*.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630.

Naihe Feng, Yi Sui, Shiyi Hou, Jesse C Cresswell, and Ga Wu. 2025. Response quality assessment for retrieval-augmented generation via conditional conformal factuality. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2832–2836.

Yichao Fu, Xuewei Wang, Yuandong Tian, and Jiawei Zhao. 2025. Deep think with confidence. *arXiv preprint arXiv:2508.15260*.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. 2024. 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)*, pages 6577–6595.

Tilmann Gneiting and Adrian E Raftery. 2007. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378.

Saumya Goswami and Siddharth Kurra. 2025. Halt-rag: A task-adaptable framework for hallucination detection with calibrated nli ensembles and abstention. *arXiv preprint arXiv:2509.07475*.

Tobias Groot and Matias Valdenegro Toro. 2024. Overconfidence is key: Verbalized uncertainty evaluation in large language and vision-language models. In *Proceedings of the 4th Workshop on Trustworthy Natural Language Processing (TrustNLP 2024)*, pages 145–171, Mexico City, Mexico. Association for Computational Linguistics.

Neha Gupta, Harikrishna Narasimhan, Wittawat Jitkrittum, Ankit Singh Rawat, Aditya Krishna Menon, and Sanjiv Kumar. 2024. Language model cascades: Token-level uncertainty and beyond. *arXiv preprint arXiv:2404.10136*.

Jiatong Han, Jannik Kossen, Muhammed Razzak, Lisa Schut, Shreshth A Malik, and Yarin Gal. 2024. Semantic entropy probes: Robust and cheap hallucination detection in LLMs. In *ICML 2024 Workshop on Foundation Models in the Wild*.

Jindong Han, Hao Liu, Jun Fang, Naiqiang Tan, and Hui Xiong. 2025. Automatic instruction data selection for large language models via uncertainty-aware influence maximization. In *Proceedings of the ACM on Web Conference 2025*, pages 4969–4979.

Cheng Peng Huang and Hao-Yuan Chen. 2025. Delta–contrastive decoding mitigates text hallucinations in large language models. *arXiv preprint arXiv:2502.05825*.

Chengsong Huang, Langlin Huang, Jixuan Leng, Jiacheng Liu, and Jiaxin Huang. 2025a. Efficient test-time scaling via self-calibration. *arXiv preprint arXiv:2503.00031*.

Haiduo Huang, Jiangcheng Song, Yadong Zhang, and Pengju Ren. 2025b. Selectkd: Selective token-weighted knowledge distillation for llms. *arXiv preprint arXiv:2510.24021*.

Zhiqi Huang, Vivek Datla, Chenyang Zhu, Alfy Samuel, Daben Liu, Anoop Kumar, and Ritesh Soni. 2025c. Confidence-based response abstinence: Improving llm trustworthiness via activation-based uncertainty estimation. In *Proceedings of the 2nd Workshop on Uncertainty-Aware NLP (UncertaiNLP 2025)*, pages 184–193.

Tomoaki Isoda. 2025. Skill-rag: Self-knowledge induced learning and filtering for retrieval-augmented generation. *arXiv preprint arXiv:2509.20377*.

Eunhye Jeong and Yong Suk Choi. 2025. Acr: Adaptive confidence re-scoring for reliable answer selection among multiple candidates. *Applied Sciences*, 15(17):9587.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong C Park. 2024. Adaptive-rag: Learning to adapt retrieval-augmented large language models through question complexity. *arXiv preprint arXiv:2403.14403*.

Ziwei Ji, Lei Yu, Yeskendir Koishekenov, Yejin Bang, Anthony Hartshorn, Alan Schelten, Cheng Zhang, Pascale Fung, and Nicola Cancedda. 2025. Calibrating verbal uncertainty as a linear feature to reduce hallucinations. *arXiv preprint arXiv:2503.14477*.

Zhengbao Jiang, Frank F Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992.

Zhengping Jiang, Anqi Liu, and Benjamin Van Durme. 2025. Conformal linguistic calibration: Trading-off between factuality and specificity. *arXiv preprint arXiv:2502.19110*.

Seongho Joo, Kyungmin Min, Jahyun Koo, and Kyomin Jung. 2025. Black-box hallucination detection via consistency under the uncertain expression. *arXiv preprint arXiv:2509.21999*.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, and 1 others. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Anant Khandelwal, Manish Gupta, and Puneet Agrawal. 2025. Cocoa: Confidence-and context-aware adaptive decoding for resolving knowledge conflicts in large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 6846–6866.

Ranganath Krishnan, Piyush Khanna, and Omesh Tickoo. 2024. Enhancing trust in large language models with uncertainty-aware fine-tuning. *arXiv preprint arXiv:2412.02904*.

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

Jixuan Leng, Chengsong Huang, Banghua Zhu, and Jiaxin Huang. 2024. Taming overconfidence in llms: Reward calibration in rlhf. *arXiv preprint arXiv:2410.09724*.

Bo Li, Tian Tian, Zhenghua Xu, Hao Cheng, Shikun Zhang, and Wei Ye. 2025a. Modeling uncertainty trends for timely retrieval in dynamic rag. *arXiv preprint arXiv:2511.09980*.

Haodong Li, Jingqi Zhang, Xiao Cheng, Peihua Mai, Haoyu Wang, and Yang Pan. 2025b. As if we've met before: Llms exhibit certainty in recognizing seen files. *arXiv preprint arXiv:2511.15192*.

Jiaqi Li, Yixuan Tang, and Yi Yang. 2025c. Know the unknown: An uncertainty-sensitive method for llm instruction tuning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 2972–2989.

Ming Li, Lichang Chen, Jiuhai Chen, Shwai He, Jiuxiang Gu, and Tianyi Zhou. 2024a. Selective reflection-tuning: Student-selected data recycling for llm instruction-tuning. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 16189–16211.

Ming Li, Yong Zhang, Shwai He, Zhitao Li, Hongyu Zhao, Jianzong Wang, Ning Cheng, and Tianyi Zhou. 2024b. Superfiltering: Weak-to-strong data filtering for fast instruction-tuning. *arXiv preprint arXiv:2402.00530*.

Ming Li, Yong Zhang, Zhitao Li, Jiuhai Chen, Lichang Chen, Ning Cheng, Jianzong Wang, Tianyi Zhou, and Jing Xiao. 2024c. From quantity to quality: Boosting llm performance with self-guided data selection for instruction tuning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7602–7635.

Pengyi Li, Matvey Skripkin, Alexander Zubrey, Andrey Kuznetsov, and Ivan Oseledets. 2025d. Confidence is all you need: Few-shot rl fine-tuning of language models. *arXiv preprint arXiv:2506.06395*.

Shuo Li, Sangdon Park, Insup Lee, and Osbert Bastani. 2024d. Traq: Trustworthy retrieval augmented question answering via conformal prediction. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3799–3821.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori B Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2023. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 12286–12312.

Yubo Li, Yidi Miao, Xueying Ding, Ramayya Krishnan, and Rema Padman. 2025e. Firm or fickle? evaluating large language models consistency in sequential interactions. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 6679–6700, Vienna, Austria. Association for Computational Linguistics.

Zixuan Li, Jing Xiong, Fanghua Ye, Chuanyang Zheng, Xun Wu, Jianqiao Lu, Zhongwei Wan, Xiaodan Liang, Chengming Li, Zhenan Sun, and 1 others. 2024e. Uncertaintyrag: Span-level uncertainty enhanced long-context modeling for retrieval-augmented generation. *arXiv preprint arXiv:2410.02719*.

Aiwei Liu, Haoping Bai, Zhiyun Lu, Yanchao Sun, Xiang Kong, Simon Wang, Jiulong Shan, Albin Madappally Jose, Xiaojiang Liu, Lijie Wen, and 1 others. 2024a. Tis-dpo: Token-level importance sampling for direct preference optimization with estimated weights. *arXiv preprint arXiv:2410.04350*.

Haotian Liu, Shuo Wang, and Hongteng Xu. 2025a. c^2gspg: Confidence-calibrated group sequence policy gradient towards self-aware reasoning. *arXiv preprint arXiv:2509.23129*.

Liangxin Liu, Xuebo Liu, Derek F Wong, Dongfang Li, Ziyi Wang, Baotian Hu, and Min Zhang. 2024b. Selectit: Selective instruction tuning for large language models via uncertainty-aware self-reflection. *arXiv preprint arXiv:2402.16705*.

Xiaoou Liu, Tiejin Chen, Longchao Da, Chacha Chen, Zhen Lin, and Hua Wei. 2025b. Uncertainty quantification and confidence calibration in large language models: A survey. *arXiv preprint arXiv:2503.15850*.

Yueyue Liu, Hongyu Zhang, Yuantian Miao, Van-Hoang Le, and Zhiqiang Li. 2024c. Optllm: Optimal assignment of queries to large language models. In *2024 IEEE International Conference on Web Services (ICWS)*, pages 788–798. IEEE.

Junjie Lu, Yuliang Liu, Chaofeng Qu, Wei Shen, Zhouhan Lin, and Min Xu. 2025. Enhancing llm reasoning via non-human-like reasoning path preference optimization. *arXiv preprint arXiv:2510.11104*.

Junyu Luo, Xiao Luo, Xiusi Chen, Zhiping Xiao, Wei Ju, and Ming Zhang. 2024. Semievol: semi-supervised fine-tuning for llm adaptation. *arXiv e-prints*, pages arXiv–2410.

Panagiotis Lymperopoulos and Vasanth Sarathy. 2025. Tools in the loop: Quantifying uncertainty of llm question answering systems that use tools. *arXiv preprint arXiv:2505.16113*.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. 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*, pages 9004–9017.

Christopher Mohri and Tatsunori Hashimoto. 2024. Language models with conformal factuality guarantees. In *International Conference on Machine Learning*, pages 36029–36047. PMLR.

Mohammad Mahdi Moradi, Hossam Amer, Sudhir Mudur, Weiwei Zhang, Yang Liu, and Walid Ahmed. 2025. Continuous self-improvement of large language models by test-time training with verifier-driven sample selection. *arXiv preprint arXiv:2505.19475*.

Diyana Muhammed, Gollam Rabby, and Sören Auer. 2025. Selfcheckagent: Zero-resource hallucination detection in generative large language models. *arXiv preprint arXiv:2502.01812*.

William Muldrew, Peter Hayes, Mingtian Zhang, and David Barber. 2024. Active preference learning for large language models. *arXiv preprint arXiv:2402.08114*.

Harikrishna Narasimhan, Wittawat Jitkrittum, Ankit Singh Rawat, Seungyeon Kim, Neha Gupta, Aditya Krishna Menon, and Sanjiv Kumar. 2024. Faster cascades via speculative decoding. *arXiv preprint arXiv:2405.19261*.

Stephen Obadinma, Xiaodan Zhu, and Hongyu Guo. 2024. Calibration attacks: A comprehensive study of adversarial attacks on model confidence. *arXiv preprint arXiv:2401.02718*.

Yijun Pan, Taiwei Shi, Jieyu Zhao, and Jiaqi W Ma. 2025. Detecting and filtering unsafe training data via data attribution. *arXiv preprint arXiv:2502.11411*.

Jinlong Pang, Na Di, Zhaowei Zhu, Jiaheng Wei, Hao Cheng, Chen Qian, and Yang Liu. 2025. Token cleaning: Fine-grained data selection for llm supervised fine-tuning. *arXiv preprint arXiv:2502.01968*.

Rhitabrat Pokharel, Yufei Tao, and Ameeta Agrawal. 2025. Capo: Confidence aware preference optimization learning for multilingual preferences. *arXiv preprint arXiv:2511.07691*.

Mihir Prabhudesai, Lili Chen, Alex Ippoliti, Katerina Fragkiadaki, Hao Liu, and Deepak Pathak. 2025. Maximizing confidence alone improves reasoning. *arXiv preprint arXiv:2505.22660*.

Ziqing Qiao, Yongheng Deng, Jiali Zeng, Dong Wang, Lai Wei, Guanbo Wang, Fandong Meng, Jie Zhou, Ju Ren, and Yaoxue Zhang. 2025. Concise: Confidence-guided compression in step-by-step efficient reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 8021–8040.

Victor Quach, Adam Fisch, Tal Schuster, Adam Yala, Jae Ho Sohn, Tommi S. Jaakkola, and Regina Barzilay. 2024. Conformal language modeling. In *The Twelfth International Conference on Learning Representations*.

Ernesto Quevedo, Jorge Yero Salazar, Rachel Koerner, Pablo Rivas, and Tomas Cerny. 2025. Detecting hallucinations in large language model generation: A token probability approach. In *Artificial Intelligence and Applications*, pages 154–173, Cham. Springer Nature Switzerland.

Stephan Rabanser, Nathalie Rauschmayr, Achin Kulshrestha, Petra Poklukar, Wittawat Jitkrittum, Sean Augenstein, Congchao Wang, and Federico Tombari. 2025. Gatekeeper: Improving model cascades through confidence tuning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

Amir Hossein Rahmati, Sanket Jantre, Weifeng Zhang, Yucheng Wang, Byung-Jun Yoon, Nathan M Urban, and Xiaoning Qian. 2025. C-lora: Contextual low-rank adaptation for uncertainty estimation in large language models. *arXiv preprint arXiv:2505.17773*.

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

Jie Ren, Yao Zhao, Tu Vu, Peter J. Liu, and Balaji Lakshminarayanan. 2023. Self-evaluation improves selective generation in large language models. In *Proceedings on "I Can't Believe It's Not Better: Failure Modes in the Age of Foundation Models" at NeurIPS 2023 Workshops*, volume 239 of *Proceedings of Machine Learning Research*, pages 49–64. PMLR.

Pouria Rouzrokh, Shahriar Faghani, Cooper U Gamble, Moein Shariatnia, and Bradley J Erickson. 2024. Conflare: conformal large language model retrieval. *arXiv preprint arXiv:2404.04287*.

Maxon Rubin-Toles, Maya Gambhir, Keshav Ramji, Aaron Roth, and Surbhi Goel. 2025. Conformal language model reasoning with coherent factuality. In *The Thirteenth International Conference on Learning Representations*.

Noveen Sachdeva, Benjamin Coleman, Wang-Cheng Kang, Jianmo Ni, Lichan Hong, Ed H Chi, James Caverlee, Julian McAuley, and Derek Zhiyuan Cheng. 2024. How to train data-efficient llms. *arXiv preprint arXiv:2402.09668*.

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

Soham Shah and Kumar Shridhar. 2025. Select-then-route : Taxonomy guided routing for LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 425–441, Suzhou (China). Association for Computational Linguistics.

Ola Shorinwa, Zhiting Mei, Justin Lidard, Allen Z Ren, and Anirudha Majumdar. 2025. A survey on uncertainty quantification of large language models: Taxonomy, open research challenges, and future directions. *ACM Computing Surveys*.

Heydar Soudani, Evangelos Kanoulas, and Faegheh Hasibi. 2025. Why uncertainty estimation methods fall short in rag: An axiomatic analysis. *arXiv preprint arXiv:2505.07459*.

Jiayuan Su, Jing Luo, Hongwei Wang, and Lu Cheng. 2024a. Api is enough: Conformal prediction for large language models without logit-access. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 979–995.

Weihang Su, Yichen Tang, Qingyao Ai, Zhijing Wu, and Yiqun Liu. 2024b. Dragin: dynamic retrieval augmented generation based on the information needs of large language models. *arXiv preprint arXiv:2403.10081*.

Xin Sun, Jianan Xie, Zhongqi Chen, Qiang Liu, Shu Wu, Yuehe Chen, Bowen Song, Zilei Wang, Weiqiang Wang, and Liang Wang. 2025. Divide-then-align: Honest alignment based on the knowledge boundary of rag. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11461–11480.

Zhongxiang Sun, Xiaoxue Zang, Kai Zheng, Yang Song, Jun Xu, Xiao Zhang, Weijie Yu, and Han Li. 2024. Redeep: Detecting hallucination in retrieval-augmented generation via mechanistic interpretability. *arXiv preprint arXiv:2410.11414*.

Shuchang Tao, Liuyi Yao, Hanxing Ding, Yuexiang Xie, Qi Cao, Fei Sun, Jinyang Gao, Huawei Shen, and Bolin Ding. 2024. When to trust llms: Aligning confidence with response quality. *arXiv preprint arXiv:2404.17287*.

Amir Taubenfeld, Tom Sheffer, Eran Ofek, Amir Feder, Ariel Goldstein, Zorik Gekhman, and Gal Yona. 2025. Confidence improves self-consistency in llms. *arXiv preprint arXiv:2502.06233*.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442.

Zailong Tian, Zhuoheng Han, Yanzhe Chen, Haozhe Xu, Xi Yang, Richeng Xuan, Houfeng Wang, and Lizi Liao. 2025. Overconfidence in llm-as-a-judge: Diagnosis and confidence-driven solution. *arXiv preprint arXiv:2508.06225*.

Benedict Aaron Tjandra, Muhammed Razzak, Jannik Kossen, Kunal Handa, and Yarin Gal. 2024. Fine-tuning large language models to appropriately abstain with semantic entropy. In *Neurips Safe Generative AI Workshop 2024*.

Sahil Tripathi, Md Tabrez Nafis, Imran Hussain, and Jiechao Gao. 2025. The confidence paradox: Can llm know when it's wrong. *arXiv preprint arXiv:2506.23464*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Yile Wang, Peng Li, Maosong Sun, and Yang Liu. 2023a. Self-knowledge guided retrieval augmentation for large language models. *arXiv preprint arXiv:2310.05002*.

Zhiruo Wang, Jun Araki, Zhengbao Jiang, Md Rizwan Parvez, and Graham Neubig. 2023b. Learning to filter context for retrieval-augmented generation. *arXiv preprint arXiv:2311.08377*.

Zihan Wang, Zihan Liang, Zhou Shao, Yufei Ma, Huangyu Dai, Ben Chen, Lingtao Mao, Chenyi Lei, Yuqing Ding, and Han Li. 2025. Infogain-rag: Boosting retrieval-augmented generation through document information gain-based reranking and filtering. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 7201–7215.

Chengwei Wei, Kee Kiat Koo, Amir Tavanaei, and Karim Bouyarmane. 2024. Confidence-aware substructure beam search (cabs): Mitigating hallucination in structured data generation with large language models. *arXiv preprint arXiv:2406.00069*.

Bingbing Wen, Jihan Yao, Shangbin Feng, Chenjun Xu, Yulia Tsvetkov, Bill Howe, and Lucy Lu Wang. 2025. Know your limits: A survey of abstention in large language models. *Transactions of the Association for Computational Linguistics*, 13:529–556.

Jiayun Wu, Jiashuo Liu, Zhiyuan Zeng, Tianyang Zhan, and Wenhao Huang. 2025. Mitigating llm hallucination via behaviorally calibrated reinforcement learning. *arXiv preprint arXiv:2512.19920*.

HuaJun Xi, Jianguo Huang, Kangdao Liu, Lei Feng, and Hongxin Wei. 2025. Does confidence calibration improve conformal prediction? *Transactions on Machine Learning Research*.

Liangru Xie, Hui Liu, Jingying Zeng, Xianfeng Tang, Yan Han, Chen Luo, Jing Huang, Zhen Li, Suhang Wang, and Qi He. 2024. A survey of calibration process for black-box llms. *arXiv preprint arXiv:2412.12767*.

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

Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, and Zhen-Hua Ling. 2024. Corrective retrieval augmented generation.

Binwei Yao, Zefan Cai, Yun-Shiuan Chuang, Shanglin Yang, Ming Jiang, Diyi Yang, and Junjie Hu. 2024. No preference left behind: Group distributional preference optimization. *arXiv preprint arXiv:2412.20299*.

Zijun Yao, Weijian Qi, Liangming Pan, Shulin Cao, Linmei Hu, Liu Weichuan, Lei Hou, and Juanzi Li. 2025. Seakr: Self-aware knowledge retrieval for adaptive retrieval augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 27022–27043.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuan-Jing Huang. 2023. Do large language models know what they don't know? In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8653–8665.

Hee Suk Yoon, Eunseop Yoon, Mark A Hasegawa-Johnson, Sungwoong Kim, and Chang D Yoo. Confpo: Exploiting policy model confidence for critical token selection in preference optimization. In *Forty-second International Conference on Machine Learning*.

Michael J. Zellinger and Matt Thomson. 2025. Rational tuning of LLM cascades via probabilistic modeling. *Transactions on Machine Learning Research*.

Yuanzhao Zhai, Han Zhang, Yu Lei, Yue Yu, Kele Xu, Dawei Feng, Bo Ding, and Huaimin Wang. 2023. Uncertainty-penalized reinforcement learning from human feedback with diverse reward lora ensembles. *arXiv preprint arXiv:2401.00243*.

Hailin Zhang, Defang Chen, and Can Wang. 2022. Confidence-aware multi-teacher knowledge distillation. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4498–4502. IEEE.

Hongxiang Zhang, Hao Chen, Muhao Chen, and Tianyi Zhang. 2025a. Active layer-contrastive decoding reduces hallucination in large language model generation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 3028–3046.

Ling Zhang, Xianliang Yang, Juwon Yu, Park Cheonyoung, Lei Song, and Jiang Bian. 2025b. Holdout-loss-based data selection for llm finetuning via in-context learning. *arXiv preprint arXiv:2510.14459*.

Tuo Zhang, Asal Mehradfar, Dimitrios Dimitriadis, and Salman Avestimehr. 2025c. Leveraging uncertainty estimation for efficient llm routing. *arXiv preprint arXiv:2502.11021*.

Qinjian Zhao, Jiaqi Wang, Zhiqiang Gao, Zhihao Dou, Belal Abuhaija, and Kaizhu Huang. 2025. Safebehavior: Simulating human-like multistage reasoning to mitigate jailbreak attacks in large language models. *arXiv preprint arXiv:2509.26345*.

Qihuang Zhong, Liang Ding, Li Shen, Juhua Liu, Bo Du, and Dacheng Tao. 2024. Revisiting knowledge distillation for autoregressive language models. *arXiv preprint arXiv:2402.11890*.

Zhanke Zhou, Xiangyu Lu, Chentao Cao, Brando Miranda, Tongliang Liu, Bo Han, and Sanmi Koyejo. 2025a. Codapo: Confidence and difficulty-adaptive policy optimization for post-training language models. In *2nd AI for Math Workshop@ ICML 2025*.

Ziang Zhou, Tianyuan Jin, Jieming Shi, and Qing Li. 2025b. Steerconf: Steering llms for confidence elicitation. *arXiv preprint arXiv:2503.02863*.

Yun Zhu, Jia-Chen Gu, Caitlin Sikora, Ho Ko, Yinxiao Liu, Chu-Cheng Lin, Lei Shu, Liangchen Luo, Lei Meng, Bang Liu, and 1 others. 2024. Accelerating inference of retrieval-augmented generation via sparse context selection. *arXiv preprint arXiv:2405.16178*.

Hanna Zubkova, Ji-Hoon Park, and Seong-Whan Lee. 2025. Sugar: Leveraging contextual confidence for smarter retrieval. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

A Notation List

Table 1 summarizes the core notation used throughout this survey; less common symbols are defined where they first appear.

Notation	Name	Description
x_i	query	the i-th query
y_i	LLM's response	response to the i-th query
c_j	candidate	the j-th candidate response of a query or a candidate reasoning trace
$\mathcal{C}$	candidate set	a set of candidate responses of a query or candidate reasoning traces
z_t	inference state	the model's inference state at step t, such as a partial reasoning trace, an intermediate generation state, or a dialogue state in multi-turn interaction
s_j, s_t	confidence score	confidence score for a candidate c_j or a reasoning step z_t
τ	threshold	confidence score threshold for inference termination or further refinement
ϵ	confidence tolerance	tolerance for changes in confidence used to decide reasoning termination
λ	contrastive weight	controls the strength of contrast between predictive views
π	inference policy	conditional distribution governing inference transitions, such as reasoning-step evolution or token generation

Table 1: Notations list