Conformal Uncertainty in LLMs: Taxonomy, Evaluation, and Open Problems

Anonymous ACL submission

Abstract

The rapid evolution of large language models (LLMs) and natural language processing (NLP) has raised growing concerns about how to quantify and communicate uncertainty across diverse tasks. Conformal prediction (CP) offers a distribution-free, model-agnostic framework for constructing uncertainty sets with finite-sample guarantees under mild assumptions, making it particularly attractive for black-box LLM deployments. Existing surveys (e.g., (Campos et al., 2024)) summarize classical CP methodology and early NLP applications, but recent progress in LLM-centric settings, including open-ended generation, reasoning, multimodal systems, and factuality, has rapidly expanded both the technical toolkit and the evaluation protocols used to validate reliability. This survey synthesizes these new developments by organizing recent CP methods for LLMs and mapping them to representative NLP-related applications, with an emphasis on how different design choices translate into practical uncertainty statements. We conclude by highlighting emerging challenges and open directions for making CP a dependable component of reliable LLM deployment.

1 Introduction

Large language models (LLMs) have rapidly reshaped modern natural language processing (NLP).This transformation is rooted in the Transformer architecture (Vaswani et al., 2017)and the pretraining paradigm that yields general-purpose language representations adaptable to diverse downstream tasks (Devlin et al., 2019; Raffel et al., 2020). Several developments have made LLMs more usable in real-world settings.Instruction tuning and reinforcement learning from human feedback (RLHF) improve instruction following and reduce certain undesirable behaviors (Ouyang et al., 2022). Prompting strategies that elicit intermediate reasoning, such as chain-of-thought prompting,often enhance performance on multi-step reasoning tasks (Wei et al., 2022).Retrieval-augmented generation (RAG) equips LLMs with access to external corpora,improving knowledge-intensive tasks and providing provenance signals (Lewis et al., 2020).

Despite these breakthroughs, uncertainty quantification remain to be questions. For instance, LLMs may produce fluent but hallucinations and can fail truthfulness benchmarks (Lin et al., 2022). Thus raises safety concerns: models can be manipulated by adversarial prompts and may generate harmful content like jailbreak attack (Zou et al., 2023); systematic red-teaming studies demonstrate that such risks persist and evolve with model scale and alignment choices (Ganguli et al., 2022). These issues are amplified under distribution shift , where raw model confidence is frequently miscalibrated.

CP offers a principled route to uncertainty quantification and risk control by transforming arbitrary scores into prediction sets with finite-sample, distribution-free guarantees under mild assumptions (Vovk et al., 2005; Angelopoulos and Bates, 2021). Early work on CP for LLMs largely targeted close-ended tasks such as multiple-choice QA, where the finite label space makes set-valued prediction and selective answering straightforward (Kumar et al., 2023). Moreover, current LLM deployments emphasize open-ended generation, black-box API access, retrieval-augmented pipelines, and non i.i.d. test conditions, which require new conformal designs. Recent directions include CP for free-form decoding (Quach et al., 2024; Deutschmann et al., 2023), logit-free/API-only uncertainty estimation (Su et al., 2024), safety-oriented conformal layers for abstention and factuality control (Abbasi-Yadkori et al., 2024; Mohri and Hashimoto, 2024), and conformal calibration for retrieval and domain shift in RAG-style systems (Rouzrokh et al., 2024; Sun et al., 2024; Lin et al., 2025).

Organization of This Survey This survey synthesizes recent progress on conformal prediction for large language models. We begin with a concise primer on CP, outlining its core principles and explaining how it provides distribution free uncertainty quantification under mild assumptions. We then review methodological advances that are especially relevant to LLM settings, including multi source CP, multivariate CP, and approaches for distribution shift and adaptive calibration (Section 2: Theoretical Conformal Prediction). In the second part, we survey representative works that apply CP across diverse NLP tasks and LLM based pipelines, highlighting common design patterns and evaluation practices (Section 3: Conformal Prediction for large language models). We conclude by distilling open challenges and outlining promising directions for future research on CP for LLMs.

2 Conformal Prediction

In this section, we will introduce the theoretical work about CP, and summarize the recent new works that may be applicable for NLP and LLMs.

Theoretical work Formalized by (Vovk et al., 2005), CP provides a rigorous framework for distribution-free uncertainty quantification with finite-sample guarantees. Recent breakthroughs in data-efficient resampling (Barber et al., 2021) and adaptive scoring (Romano et al., 2019, 2020) have matured the paradigm, enabling prediction regions to scale with instance-level difficulty. This theoretical foundation has recently expanded to address the non-exchangeability inherent in NLP via Wasserstein-regularized distribution shifts (Xu et al., 2025a) and generalized risk management (Xu et al., 2025b).

General framework for conformal prediction. CP transforms a point predictor $\hat{f}$ into a set-valued output $\mathcal{C}_\alpha(x)$ that guarantees $P(Y \in \mathcal{C}_\alpha(x)) \geq 1-\alpha$ under exchangeability. In the split conformal setting (Papadopoulos et al., 2002; Lei et al., 2018), calibration is performed on a held-out set $\mathcal{D}_{cal} = \{(X_i, Y_i)\}_{i=1}^n$ by defining a task-specific nonconformity score $s(x, y)$. Common instantiations of $s(x, y)$ include cumulative softmax mass for classification (Romano et al., 2020). The calibration threshold $\hat{q}$ is computed as the $\frac{\lceil (n+1)(1-\alpha) \rceil}{n}$-th empirical quantile of the scores $\{s(X_i, Y_i)\}_{i=1}^n$:

$$\hat{q} = \inf\left\{ q : \frac{|\{i : s_i \leq q\}|}{n} \geq \frac{\lceil (n+1)(1-\alpha) \rceil}{n} \right\}.$$

The resulting prediction set is defined as $\mathcal{C}_\alpha(x) = \{y : s(x, y) \leq \hat{q}\}$. While this baseline assumes exchangeability, recent extensions such as Wasserstein-regularized CP (Xu et al., 2025a) further relax these constraints to handle complex distribution shifts.

Conformal risk control Conformal Risk Control (CRC) formalizes the calibration objective by bounding a user-specified loss $L(\cdot)$, such as factual error or safety violation, to ensure that the population risk $R(\lambda) = \mathbb{E}[L(\mathcal{A}_\lambda(X), Y)]$ remains below a threshold α. This framework is uniquely effective for LLM alignment, where rare but severe failures necessitate rigorous control. For example, (Chen et al., 2025) introduced tail-risk formulations for black-box models, providing finite-sample guarantees for distortion risk measures targeting extreme losses. To improve operational efficiency, recent methods address LLM sampling variability via randomized and bootstrapped variants, which stabilize thresholds and reduce calibration overhead (Pang et al., 2025).

Beyond post-generation filtering, CRC enables modular decision-making within LLM pipelines. In Retrieval-Augmented Generation (RAG), conformalized retrieval dynamically scales context sets to guarantee the inclusion of ground-truth evidence, thereby mitigating hallucinations caused by context omission (Li et al., 2024a). Additionally, selective frameworks like COIN and SConU combine CRC guarantees with FDR-style risk constraints and outlier detection to prune unreliable outputs (Wang et al., 2025c,d). These advancements converge toward "conformal arbitrage," a system-level paradigm for routing queries between primary models and conservative guardians to optimize the trade-off between helpfulness and provable risk (Overman and Bayati, 2025; Xu et al., 2025b).

Distribution shift and adaptive conformal prediction. Classical conformal prediction relies on the exchangeability assumption, which is frequently violated in real-world LLM deployments due to distribution shifts across domains, tasks, and prompts. To maintain coverage validity under such shifts, adaptive methods replace global thresholds with test-point-specific quantiles derived from weighted calibration scores S_i:

$$\hat{q}_{1-\alpha}(x) = \inf\left\{ q : \sum_{i \in \mathcal{I}_{\text{cal}}} w_i(x)\, \mathbf{1}\{S_i \leq q\} \geq 1-\alpha \right\}$$

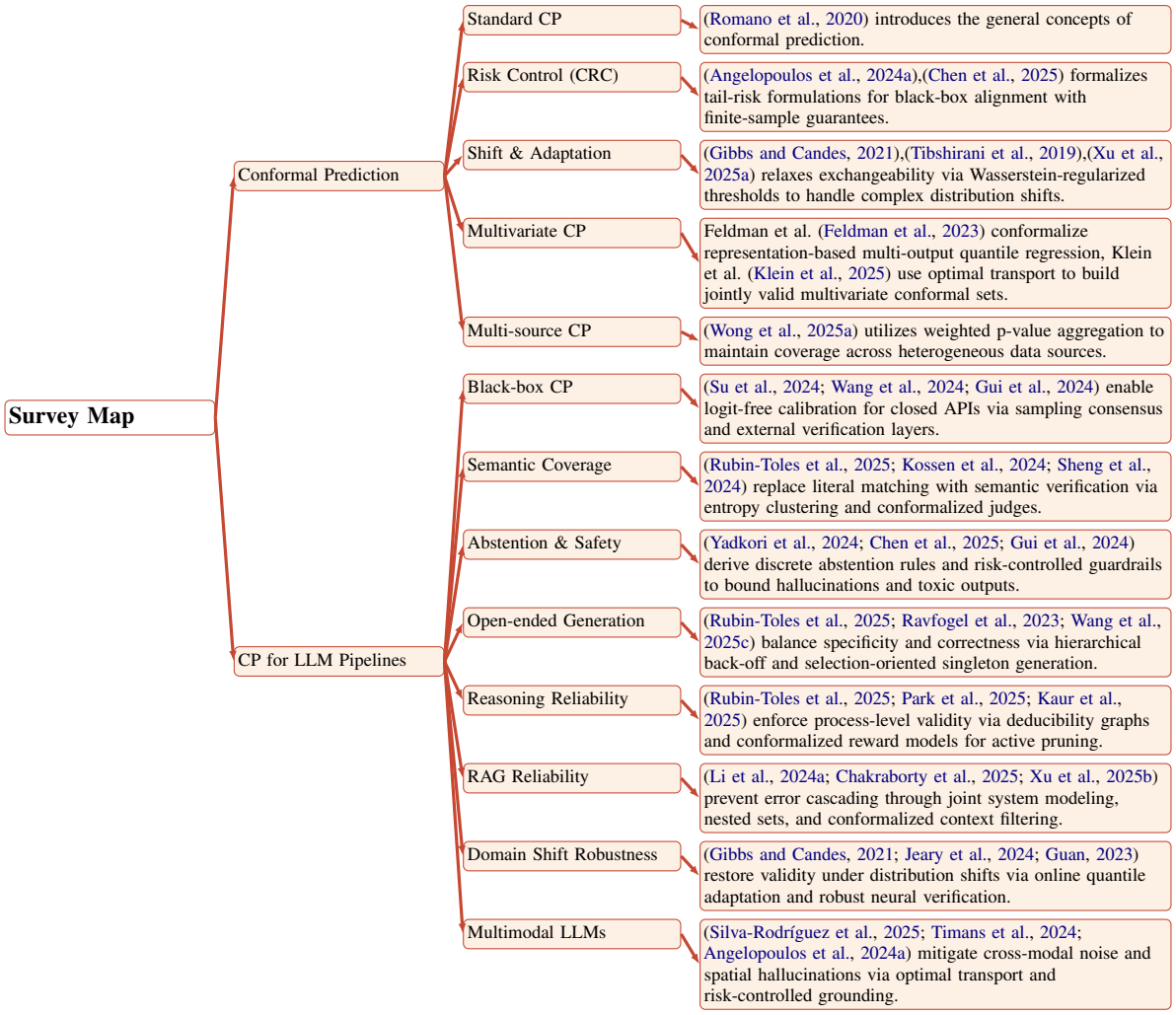

Figure 1: Compact taxonomy of conformal prediction foundations (Section 2) and LLM pipeline applications (Section 3).

where non-negative weights $w_i(x)$ prioritize calibration points most relevant to the test input x. This paradigm is realized through subpopulation reweighting and domain-aware wrappers that leverage task-specific information to mitigate shift-induced under-coverage (Wang et al., 2025a; Lin et al., 2025). A specialized extension for LLMs is In-Context Conformal Prediction (ICCP), which enables local calibration by using few-shot demonstrations within the prompt as a dynamic calibration set, adapting to task distributions without requiring external data (Deng et al., 2025). Beyond reweighting, recent research explores adaptive coverage policies $\alpha(x)$ to optimize efficiency for heterogeneous queries (Gauthier et al., 2025), alongside distribution-informed online updates for rapid recalibration in non-stationary streaming environments (Jun and Ohn, 2025; Hu et al., 2025). These data-dependent thresholds and online adjustments establish a robust foundation for uncertainty quantification in modular LLM pipelines (Zhang et al., 2025; Badkul and Xie, 2025).

Multivariate conformal prediction. Although CP typically relies on scalar nonconformity scores, its extension to vector-valued outputs is non-trivial due to the absence of a canonical ordering in higher dimensions. One primary approach involves learning tractable representations of multivariate responses to calibrate flexible base regions, enabling expressive, non-rectangular predictive sets that maintain exact marginal coverage (Feldman et al., 2023). Alternatively, optimal transport (OT) provides a principled framework for defining multivariate ranks and quantiles to induce an ordering for vector-valued scores. This allows for the construction of valid prediction sets in multidimensional regression via OT-based ranking protocols (Klein et al., 2025). Similar methodologies uti-

lize Monge–Kantorovich vector ranks to develop flexible, potentially non-convex conformal regions for multi-output regression and multiclass classification, with certain variants specifically targeting stronger conditional guarantees (Thurin et al., 2025).

Multi-source conformal prediction and online conformal prediction. LLM deployments often utilize calibration data from multiple sources, where naive pooling risks violating exchangeability and degrading coverage validity. This challenge is addressed through source-aware conformalization, which calibrates within individual sources and aggregates evidence via weighted p-values to ensure robustness against variations in the source mixture (Wong et al., 2025a). When distribution shifts stem from changing proportions of latent groups, adaptive procedures employing sample reweighting or test-dependent thresholds can maintain validity under unknown subpopulation shifts (Wang et al., 2025a). In federated settings, Byzantine-robust methodologies preserve distribution-free guarantees even when a subset of participants reports manipulated calibration statistics (Kang et al., 2024). For non-stationary streaming regimes, such as evolving prompt distributions, online conformal prediction updates thresholds sequentially. The use of decaying step sizes in these environments provides strong worst-case guarantees while improving the stability of time-local coverage (Angelopoulos et al., 2024b).

3 Conformal Prediction for Large Language Models

In this section, we discuss how CP techniques are deployed in LLMs settings, especially under closed-source APIs, safety and multi-stage pipelines (e.g., retrieval, generation, and verification). We review recent adaptations of CP for semantic correctness of black-box outputs, abstention and conformal risk control, open-ended generation and reasoning, RAG and context selection, robustness under domain shift, and extensions to multimodal LLMs.

Why conformal prediction for large language models The deployment of modern LLMs via closed-source interfaces decouples model utility from internal state visibility, rendering traditional white-box calibration methods that rely on log-probabilities largely obsolete (Su et al., 2024; Wang et al., 2024). This transition aligns with the emerging Model Science framework, which advocates for systematic verification and control of model behavior rather than traditional data-centric evaluation (Biecek and Samek, 2025). Within this paradigm, CP functions as a formal verification mapping V that utilizes a model $\hat{f}$ and a calibration set $\mathcal{D}_{cal}$ to construct a guaranteed uncertainty region:

$$V(\hat{f}, \mathcal{D}_{cal}, \epsilon) \to \Gamma_\epsilon(X) \quad \text{s.t.} \quad P(Y \notin \Gamma_\epsilon(X)) \leq \epsilon$$

where ϵ denotes a predefined error budget. This statistical protocol ensures that heuristic black-box generators are transformed into reliable system components with provable boundaries.

The adoption of CP is further necessitated by the inherent fallibility of LLM self-assessment. While models exhibit emergent capacities to estimate their own correctness via internal heuristics (Kadavath et al., 2022), these subjective measures lack the objective rigor required for high-stakes auditing. The absence of an external verification layer often results in uncalibrated overconfidence, a gap addressed by CP through an objective statistical layer that maps raw outputs to valid prediction sets or formal abstention rules (Gui et al., 2024). This mechanism effectively decouples system reliability from the model's internal biases.

Correctness coverage for black-box LLM outputs A primary methodological challenge in applying CP to LLMs involves the transition from categorical labels to structured semantic objects such as strings and multi-step rationales. In this context, the conventional definition of coverage must be re-anchored from literal matching to a semantic verification protocol (Rubin-Toles et al., 2025). The central question shifts from simple error counting to: How can we design invariant non-conformity scores that capture latent semantic truth rather than surface-level token overlap?

To address this, researchers have moved beyond white-box logits toward scores grounded in semantic faithfulness and logical consistency. For instance, (Su et al., 2024) formulates non-conformity measures using fine-grained uncertainty notions such as semantic similarity clusters. This perspective aligns with the broader Model Science paradigm (Biecek and Samek, 2025), which advocates for incorporating an explicit verification layer V into the generative pipeline. Formally, a robust score $s(X, Y)$ can be generalized as an alignment function:

$$s(X, Y) = 1 - \text{Score}(\Phi(Y), \Psi(X))$$

where $\Phi(Y)$ represents the semantic representation of the candidate and $\Psi(X)$ denotes the verified grounding evidence, which can be derived from internal semantic entropy clusters (Kossen et al., 2024) or external evaluative agents (Sheng et al., 2024).

Furthermore, this framework has expanded to the control of reasoning trajectories. The emergence of conformalized judges allows for the quantification of uncertainty in rating-based assessments, ensuring that LLM-as-a-judge frameworks provide statistically grounded prediction intervals rather than uncalibrated scores (Sheng et al., 2024). By integrating these measures across deducibility graphs, the field is transitioning toward a state where every LLM output is accompanied by a formal certificate of semantic validity.

From prediction sets to abstention and risk control While CP traditionally yields set-valued outputs, practical deployment within LLM ecosystems necessitates a transition toward discrete decision-making protocols such as answer deferral and selective abstention. Rather than presenting users with a raw set of potential candidates, CP provides a rigorous framework for deriving abstention policies that maintain explicit error control. A prominent application of this paradigm is the mitigation of factual hallucinations, where the system is calibrated to withhold a response, effectively stating "I do not know", whenever the uncertainty of the generated content exceeds a predefined safety threshold (Yadkori et al., 2024). By optimizing the trade-off between model utility and factual integrity, such frameworks bound the hallucination rate while maximizing the density of informative responses for answerable queries.

The scope of this control has recently evolved beyond simple error rates toward the management of complex risk functionals. Recent research into conformal tail-risk control addressing LLM alignment targets specific safety-critical distortions, such as toxic or offensive outputs, through the use of weighted quantiles (Chen et al., 2025). This approach introduces a lightweight calibration layer that remains compatible with black-box scoring systems, ensuring that even rare but catastrophic failures are statistically bounded. This shift toward multi-objective risk control allows for the simultaneous bounding of disparate risks, including factual inaccuracy and social bias, through a unified statistical interface (Angelopoulos et al., 2024a).

The emerging paradigm of conformal alignment represents a fundamental bridge between training objectives and downstream inference guarantees (Gui et al., 2024). Unlike post-hoc calibration methods, this approach suggests that models can be fine-tuned to yield probability distributions that are inherently more "conformalizable." By optimizing for distributions that produce smaller and more precise prediction sets, conformal alignment enhances the efficiency of abstention policies, ultimately positioning CP not merely as a diagnostic tool but as an integral component of the algorithmic governance and safety alignment pipeline.

Although jailbreak attacks have been extensively studied, two safety gaps persist: (i) defenses often fail to generalize across heterogeneous attack families, and (ii) guardrails are brittle to novel, unseen attack types (Wang et al., 2025b). Motivated by this, we view jailbreak as a subpopulation shift problem indexed by attack type. Let $\mathcal{A} = \{a_1, a_2, \ldots\}$ denote known attack types, let $p \in \mathcal{P}$ be an input prompt, and let $X = \phi(p) \in \mathcal{X}$ be a prompt representation (e.g., last-token embedding or a hidden-state summary). For a victim model f and an LLM-based judge J, define the judged safety outcome $Z := J(f(p)) \in \mathcal{Z}$ (e.g., $\mathcal{Z} = \{0, 1\}$ for jailbreak success/failure or a graded safety score). Our goal is to construct a conformal prediction set $\widehat{C}(X) \subseteq \mathcal{Z}$ with miscoverage level $\alpha \in (0, 1)$ that is (a) attack-conditional over seen families, i.e., $\mathbb{P}\Big(Z \in \widehat{C}(\phi(p)) \,\Big|\, p \in a_i\Big) \geq 1 - \alpha$ for all $a_i \in \mathcal{A}$, and (b) robust/adaptive to emerging families $a^\star \notin \mathcal{A}$, i.e., $\mathbb{P}\Big(Z \in \widehat{C}(\phi(p)) \,\Big|\, p \in a^\star\Big) \geq 1 - \alpha$. This formulation directly targets both failure modes above by enabling multi-source calibration across known attacks and principled adaptation when new jailbreak types appear.

Open-ended generation with conformal guarantees Adapting Conformal Prediction (CP) to open-ended generation requires a transition from static label classification toward the dynamic calibration of autoregressive trajectories. This process is constrained by the combinatorial complexity of the output space, necessitating methods that treat generation as a path within a structured semantic hierarchy rather than a sequence of independent tokens.

A primary mechanism in this domain is hierarchical back-off, which operationalizes the trade-off

between semantic specificity and statistical correctness (Biecek and Samek, 2025). When the non-conformity score of a specific claim exceeds the calibrated threshold, the system adaptively reverts to a more general yet provably correct abstraction within a predefined taxonomy (Rubin-Toles et al., 2025). This protocol ensures that LLMs remain truthful by design by modulating informativeness to match internal confidence levels. Formally, given a taxonomy $\mathcal{T}$, the system selects the most specific node $y \in \mathcal{T}$ such that the safety guarantee $\mathbb{P}(Y^* \in \text{descendants}(y)) \geq 1 - \alpha$ is preserved.

Beyond abstraction, recent advancements integrate CP into the decoding process to enable statistically principled search pruning. Unlike heuristic-based nucleus sampling, conformal decoding strategies (Ravfogel et al., 2023) dynamically adjust the candidate vocabulary $\mathcal{V}_t$ at each timestep t:

$$\mathcal{V}_t = \{w \in \Sigma \mid \sum_{j \in \text{sorted}(\Sigma)} \pi(w_j | x, y_{<t}) \leq \hat{\lambda}\}$$

where $\hat{\lambda}$ is the calibrated threshold. This approach is critical in multi-step reasoning where error propagation leads to logical drift. By imposing invariant safety bounds across the reasoning chain, these frameworks utilize deducibility structures to terminate fallacious trajectories before they culminate in hallucinated conclusions (Rubin-Toles et al., 2025).

Selection-oriented conformalization (Wang et al., 2025c) further addresses the requirement of generating high-quality singleton responses under rigorous risk control (Biecek and Samek, 2025). By evaluating candidates against user-specified loss functions L, these methods filter responses that violate stipulated risk bounds, ensuring $\mathbb{E}[L(\hat{y}, y)] \leq \alpha$. In black-box scenarios where internal logits are inaccessible, semantic self-consistency serves as a proxy for uncertainty, enabling the derivation of valid bounds through the aggregation of sampled consensus (Wang et al., 2024; Su et al., 2024).

Reasoning and factual coherence Reasoning tasks possess inherent structural dependencies where correctness is contingent upon the integrity of multi-step derivations, and factual errors often exhibit strong correlations across sequential claims. To address these dependencies, research has pivoted from outcome-level validation to process-level guarantees, treating the entire reasoning trace as a stochastically constrained trajectory. A pivotal advancement is the coherent factuality framework (Rubin-Toles et al., 2025), which utilizes deducibility graphs to enforce statistical bounds that respect logical dependencies. Formally, for a reasoning path $\mathcal{P} = \{s_1, s_2, \ldots, s_k\}$, the system ensures that the joint coverage of the entire derivation remains valid:

$$\mathbb{P}\left(\bigcap_{i=1}^{k}(s_i \in \mathcal{C}_{true})\right) \geq 1 - \alpha$$

By calibrating the premises rather than just the conclusion, this approach ensures that any derived statement inherits a formal certificate of validity from its logical antecedents.

Crucially, recent methodologies have operationalized these statistical signals to transition from passive monitoring to active search control. In the context of multi-path reasoning (e.g., Tree of Thoughts), conformal scores now function as rigorous pruning criteria. Intermediate steps violating local coverage thresholds trigger immediate backtracking, thereby dynamically reallocating compute resources to branches with valid partial proofs. This mechanism is exemplified by the CAL-PRM framework (Park et al., 2025), which conformalizes Process Reward Models (PRMs) to enable instance-adaptive search scaling. The pruning decision at step t is governed by a calibrated threshold $\hat{\lambda}_t$:

$$\text{Prune}(s_t) = \mathbb{I}(\text{Score}_{PRM}(s_t) < \hat{\lambda}_t)$$

where $\hat{\lambda}_t$ is determined to maintain a global safety envelope across long-horizon planning. Further extending this rigor to structured logic, neuro-symbolic integrations bridge fluid neural generation with symbolic scaffolds. For instance, by applying CP to the selection of Answer Set Programming (ASP) rules, generated logical proofs can maintain verified coverage even when the underlying solver operates as a black box (Kaur et al., 2025). This synergy enables the deployment of LLMs in environments requiring both generative flexibility and the absolute rigor of formal logic.

RAG, retrieval uncertainty, and principled context engineering While RAG is a primary technique for grounding LLM outputs in external evidence, it introduces a complex hierarchy of uncertainty where retrieval failures often lead to *error cascading* in the generation phase (Li et al., 2024a). Recent advancements in Conformal Prediction (CP) address this by providing a unified framework for end-to-end correctness guarantees.

Specifically, the TRAQ framework (Li et al., 2024) models the retriever and generator as a joint system, constructing nested prediction sets to ensure the presence of semantically correct answers with high probability (Li et al., 2024a). Complementing this, CONFLARE (Rouzrokh et al., 2024) quantifies retrieval uncertainty to prevent noise propagation, while recent work by Wong et al. (2025) extends this to multi-source environments by using weighted p-values to improve coverage in aggregated evidence sets (Wong et al., 2025a).

A significant frontier involves principled context engineering, where CP serves as a statistical filter for context management. Chakraborty et al. (2025) demonstrate that conformalizing relevance scores allows for the pruning of irrelevant snippets while strictly bounding the probability of discarding critical information (Chakraborty et al., 2025). This is further generalized via selective conformal risk control (Xu et al., 2025b), which adopts a "select-then-calibrate" paradigm to manage the trade-off between inference efficiency and the risk of information loss (Xu et al., 2025b). As a final safeguard, conformal abstention mechanisms (Abbasi-Yadkori et al., 2024) provide a principled fallback, allowing the model to refrain from answering when the calibrated hallucination risk exceeds a pre-defined threshold. Collectively, these methods transition RAG from a heuristic integration of search and generation into a statistically grounded framework with provable reliability.

While RAG is a primary technique for grounding LLM outputs in external evidence, it introduces a complex hierarchy of uncertainty where retrieval failures often lead to error cascading in the generation phase (Li et al., 2024a). Recent advancements in CP address this by providing a unified framework for end-to-end correctness guarantees. Specifically, the TRAQ framework (Li et al., 2024) models the retriever and generator as a joint system, constructing nested prediction sets to ensure the presence of semantically correct answers with high probability (Li et al., 2024a). Complementing this, CONFLARE (Rouzrokh et al., 2024) quantifies retrieval uncertainty to prevent noise propagation, while recent work by Wong et al. (2025) extends this to multi-source environments by using weighted p-values to improve coverage in aggregated evidence sets (Rouzrokh et al., 2024; Wong et al., 2025a).

Domain shift and deployment robustness A fundamental challenge in transitioning LLMs to dynamic production environments is the potential violation of the exchangeability assumption. In real-world deployments, covariate shifts and adversarial perturbations often lead to coverage collapse. To restore statistical validity, recent research has moved toward Adaptive Conformal Prediction (ACP), which updates the quantile $\hat{q}_t$ online based on historical coverage errors (Gibbs and Candes, 2021). However, to provide guarantees against worst-case shifts, the Verifiably Robust Conformal Prediction (VRCP) framework (Jeary et al., 2024) integrates neural network verification algorithms to bound the impact of perturbations on conformity scores. For a given perturbation set $\mathcal{B}_\epsilon(x)$, VRCP constructs robust prediction sets by calibrating against the upper bound of the non-conformity scores:

$$\mathcal{C}_{rob}(x) = \{y \in \mathcal{Y} \mid \max_{x' \in \mathcal{B}_\epsilon(x)} E(x', y) \leq \hat{q}\}$$

This approach ensures that coverage is maintained even under ℓ_p-bounded adversarial attacks, effectively bridging the gap between statistical calibration and formal safety verification. By combining these robust intervals with localized calibration (Guan, 2023), LLMs can maintain rigorous integrity in non-stationary and potentially adversarial environments.

Conformal Prediction for multimodel large language models The expansion of LLMs into multimodal domains (MLLMs) introduces heterogeneous uncertainty, compounding visual perception noise with cross-modal alignment errors. Standard calibration often fails in these high-dimensional settings, particularly under zero-shot shifts. (Silva-Rodríguez et al., 2025)Silva-Rodríguez et al. (2025) addressed the reliability of foundational vision-language models by proposing a transductive conformal framework, Conf-OT. By leveraging optimal transport to align visual and textual embeddings on the fly, this method rectifies the distributional mismatch. The non-conformity score is defined by the Wasserstein distance $W(\cdot, \cdot)$ between the test sample and the calibrated clusters:

$$s_i = \inf_{\gamma \in \Gamma(P_{cal}, P_{test})} \int \|z_{vis} - z_{txt}\|^2 d\gamma$$

This alignment ensures valid coverage for unseen classes without supervised fine-tuning (Silva-Rodríguez et al., 2025).

Beyond semantic classification, safety-critical applications require precise spatial grounding. (Timans et al., 2024) introduced a two-step conformal mechanism that first constructs a prediction set for the object category and subsequently propagates this uncertainty into the localization head. This results in adaptive prediction regions $\mathcal{R}(x)$ that guarantee the inclusion of the ground-truth bounding box B^* given a correct classification:

$$\mathbb{P}(B^* \subseteq \mathcal{R}(X) \mid Y \in \mathcal{C}_{class}) \geq 1 - \alpha$$

This effectively mitigates "spatial hallucinations" where models confidently localize misclassified entities (Timans et al., 2024).

For open-ended generative tasks like Visual Question Answering , (Angelopoulos et al., 2024a) generalized the conformal paradigm via Conformal Risk Control . Instead of bounding the error probability, CRC theoretically bounds the expected value of monotonic loss functions L (e.g., semantic dissimilarity or 1-IoU):

$$\mathbb{E}[L(\mathcal{C}_{\hat{\lambda}}(X), Y)] \leq \alpha \tag{1}$$

This allows MLLMs to dynamically adjust output set sizes to maintain a user-specified limit on factual error rates, providing a rigorous safety layer for complex vision-language reasoning (Angelopoulos et al., 2024a).

Finally, for open-ended generative tasks like Visual Question Answering (VQA), where discrete prediction sets are infeasible, (Angelopoulos et al., 2024a) generalized the conformal paradigm via Conformal Risk Control (CRC). Instead of bounding the probability of error, CRC theoretically bounds the expected value of arbitrary monotonic loss functions (e.g., semantic dissimilarity or 1-IoU). This allows MLLMs to dynamically adjust output set sizes to maintain a user-specified limit on factual error rates, providing a rigorous safety layer for complex vision-language reasoning (Angelopoulos et al., 2024a).

4 Future Work

Moving forward, the confluence of CP and LLMs suggests several critical trajectories that transcend the current static evaluation paradigms summarized in this work. A primary theoretical frontier involves the relaxation of the exchangeability assumption (Tibshirani et al., 2019), which, while foundational to classical CP, is increasingly strained by the non-stationary and adversarial nature of real-world LLM deployments. Consequently, there is a compelling need for adaptive, online calibration frameworks (Gibbs and Candes, 2021; Zaffran et al., 2022) capable of maintaining statistical rigor under continuous distribution shifts, as well as the need of LLMs for multivariate and multi-source CP.

Building on the semantic-coverage taxonomy in Section 3, future work should move beyond lexical and NLI-based nonconformity scores toward logic-aware notions of correctness, especially for structured reasoning where functional validity matters (e.g., mathematical correctness and knowledge-graph consistency) (Shahrokhi et al., 2025b; Ni et al., 2024). Moreover, Figure 1 suggests that impact will depend on integrating CP into inference rather than treating it as a post-hoc wrapper: conformal-aware decoding and test-time control can enforce uncertainty constraints while managing compute, including asynchronous test-time scaling (Xiong et al., 2025) and confidence-controlled early exits (Akgül et al., 2025).

Figure 2 further summarizes a complementary set of open questions for conformalized LLM systems. Beyond the method-and-application taxonomy, it highlights both foundational CP frontiers (e.g., improved conditional coverage, scalable resampling, and robustness under extreme shifts) and pipeline-level challenges specific to LLM deployment (e.g., long-horizon reasoning and planning, end-to-end RAG optimization, cross-modal alignment, and adversarial defense). These directions emphasize that progress will require advancing CP theory alongside system-aware designs that account for the multi-stage, non-stationary, and safety-critical nature of real-world LLM pipelines.

5 Conclusion

In this survey, we synthesize recent advances in conformal prediction (CP) for large language models and organize the space through the taxonomy in Figure 1, covering theoretical foundations, semantic scoring, and downstream applications. A central takeaway is the shift from lexical heuristics to semantic correctness objectives with distribution-free guarantees, enabling more reliable deployment in settings such as RAG and long-horizon reasoning. We conclude by highlighting open challenges in efficiency and adaptive calibration, and we hope this survey provides a practical roadmap for future work on trustworthy LLM systems.

References

Yasin Abbasi-Yadkori, Ilja Kuzborskij, David Stutz, András György, Adam Fisch, Arnaud Doucet, Iuliya Beloshapka, Wei-Hung Weng, Yao-Yuan Yang, Csaba Szepesvári, Ali Taylan Cemgil, and Nenad Tomasev. 2024. Mitigating LLM hallucinations via conformal abstention. *arXiv preprint arXiv:2405.01563*.

Ömer Faruk Akgül, Yusuf Hakan Kalaycı, Rajgopal Kannan, Willie Neiswanger, and Viktor Prasanna. 2025. Lynx: Learning dynamic exits for confidence-controlled reasoning. *arXiv preprint arXiv:2512.05325*.

Anastasios N. Angelopoulos and Stephen Bates. 2021. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*.

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

Anastasios Nikolas Angelopoulos, Rina Barber, and Stephen Bates. 2024b. Online conformal prediction with decaying step sizes. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 1616–1630. PMLR.

Amitesh Badkul and Lei Xie. 2025. Adaptive individual uncertainty under out-of-distribution shift with expert-routed conformal prediction. *Preprint*, arXiv:2510.15233.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. 2021. Predictive inference with the jackknife+. *The Annals of Statistics*, 49(1):486–507.

Przemyslaw Biecek and Wojciech Samek. 2025. Model science: getting serious about verification, explanation and control of ai systems. *arXiv preprint arXiv:2508.20040*. Accepted at the 28th European Conference on Artificial Intelligence (ECAI) 2025.

Margarida Campos, António Farinhas, Chrysoula Zerva, Mário A. T. Figueiredo, and André F. T. Martins. 2024. Conformal prediction for natural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 12:1497–1516.

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

Catherine Yu-Chi Chen, Jingyan Shen, Zhun Deng, and Lihua Lei. 2025. Conformal tail risk control for large language model alignment. *Preprint*, arXiv:2502.20285.

Weicao Deng and 1 others. 2025. Optimizing in-context learning for efficient full conformal prediction. *arXiv preprint arXiv:2509.01840*.

Nicolas Deutschmann, Marvin Alberts, and María Rodríguez Martínez. 2023. Conformal autoregressive generation: Beam search with coverage guarantees. *arXiv preprint arXiv:2309.03797*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Shai Feldman, Stephen Bates, and Yaniv Romano. 2023. Calibrated multiple-output quantile regression with representation learning. *Journal of Machine Learning Research*, 24(24):1–48.

Deep Ganguli, Amanda Askell, Nicholas Schiefer, Thomas I. Liao, Anna Chen, Adam Goldstein, Tom Henighan, Saurav Kadavath, Andy Jones, Jackson Kernion, Ben Li, Liane Lovitt, Neel Nanda, Catherine Olsson, Joe Thornton, Maria Tsimpoukelli, John Hewitt, Tom Conerly, Jared Kaplan, and 3 others. 2022. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*.

Etienne Gauthier, Francis Bach, and Michael I. Jordan. 2025. Adaptive coverage policies in conformal prediction. *Preprint*, arXiv:2510.04318.

Isaac Gibbs and Emmanuel Candes. 2021. Adaptive conformal inference under distribution shift. In *Advances in Neural Information Processing Systems*, volume 34, pages 1660–1672.

Isaac Gibbs and Emmanuel J. Candès. 2021. Adaptive conformal inference under distribution shift. *Preprint*, arXiv:2106.00170.

Leying Guan. 2023. Localized conformal prediction. *Biometrika*, 110(4):893–910. ArXiv preprint arXiv:2106.08460.

Yu Gui, Ying Jin, and Zhimei Ren. 2024. Conformal alignment: Knowing when to trust foundation models with guarantees. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Dongjian Hu, Junxi Wu, Shu-Tao Xia, and Changliang Zou. 2025. Distribution-informed online conformal prediction. *Preprint*, arXiv:2512.07770.

Linus Jeary, Nicola Paoletti, Tom Kuipers, and Mehran Hosseini. 2024. Verifiably robust conformal prediction. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Jungbin Jun and Ilsang Ohn. 2025. Online conformal inference with retrospective adjustment for faster adaptation to distribution shift. *Preprint*, arXiv:2511.04275.

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*.

Mintong Kang, Zhen Lin, Jimeng Sun, Cao Xiao, and Bo Li. 2024. Certifiably Byzantine-robust federated conformal prediction. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 23022–23057. PMLR.

Navdeep Kaur, Lachlan McPheat, Alessandra Russo, Anthony G Cohn, and Pranava Madhyastha. 2025. An empirical study of conformal prediction in LLM with ASP scaffolds for robust reasoning. *arXiv preprint arXiv:2503.05439*.

Michal Klein, Louis Béthune, Eugene Ndiaye, and Marco Cuturi. 2025. Multivariate conformal prediction using Optimal Transport.

Jannik Kossen, Jiatong Han, Muhammed Razzak, Lisa Schut, Shreshth Malik, and Yarin Gal. 2024. Semantic entropy probes: Robust and cheap hallucination detection in llms. *Preprint*, arXiv:2406.15927.

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*.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. 2018. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*.

Shuo Li, Sangdon Park, Insup Lee, and Osbert Bastani. 2024a. 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.

Shuo Li, Sangdon Park, Insup Lee, and Osbert Bastani. 2024b. 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. Association for Computational Linguistics.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Zhexiao Lin, Yuanyuan Li, Neeraj Sarna, Yuanyuan Gao, and Michael von Gablenz. 2025. Domain-shift-aware conformal prediction for large language models. *arXiv preprint arXiv:2510.05566*.

Christopher Mohri and Tatsunori Hashimoto. 2024. Language models with conformal factuality guarantees. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 36029–36047. PMLR.

Bo Ni, Yu Wang, Lu Cheng, Erik Blasch, and Tyler Derr. 2024. Towards trustworthy knowledge graph reasoning: An uncertainty aware perspective. *arXiv preprint arXiv:2410.08985*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

William Overman and Mohsen Bayati. 2025. Conformal arbitrage: Risk-controlled balancing of competing objectives in language models. *Preprint*, arXiv:2506.00911.

Lingyou Pang, Lei Huang, Jianyu Lin, Tianyu Wang, Alexander Aue, and Carey E. Priebe. 2025. Taming variability: Randomized and bootstrapped conformal risk control for LLMs. *Preprint*, arXiv:2509.23007.

Harris Papadopoulos, Kostas Proedrou, Vladimir Vovk, and Alexander Gammerman. 2002. Inductive confidence machines for regression. In *Machine Learning: ECML 2002*, volume 2430 of *Lecture Notes in Computer Science*, pages 345–356. Springer.

Young-Jin Park, Kristjan Greenewald, Kaveh Alim, Hao Wang, and Navid Azizan. 2025. Know what you don't know: Uncertainty calibration of process reward models. *arXiv preprint arXiv:2506.09338*. Accepted at NeurIPS 2025.

Victor Quach, Adam Fisch, Tal Schuster, Adam Yala, Jae Ho Sohn, Tommi S. Jaakkola, and Regina Barzilay. 2024. Conformal language modeling. *arXiv preprint arXiv:2306.10193*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Shauli Ravfogel, Yoav Goldberg, and Jacob Goldberger. 2023. Conformal nucleus sampling. *Preprint*, arXiv:2305.02633.

Yaniv Romano, Evan Patterson, and Emmanuel J. Candès. 2019. Conformalized quantile regression. In *Advances in Neural Information Processing Systems*, volume 32, pages 3538–3548. Curran Associates, Inc.

Yaniv Romano, Matteo Sesia, and Emmanuel J. Candès. 2020. Classification with valid and adaptive coverage. In *Advances in Neural Information Processing Systems*, volume 33, pages 3581–3591.

Parsa Rouzrokh, Cameron Blake, Zongyi Liang, Sina Honari, and Yi Luan. 2024. Conflare: Conformal large language model retrieval. *arXiv preprint*.

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

Hooman Shahrokhi, Devjeet Raj Roy, Yan Yan, Venera Arnaoudova, and Jana Doppa. 2025a. Conformal prediction sets for deep generative models via reduction to conformal regression. In *Proceedings of the Forty-first Conference on Uncertainty in Artificial Intelligence*, volume 286 of *Proceedings of Machine Learning Research*, pages 3718–3748. PMLR.

Hooman Shahrokhi, Devjeet Raj Roy, Yan Yan, Venera Arnaoudova, and Janaradhan Rao Doppa. 2025b. Conformal prediction sets for deep generative models via reduction to conformal regression. *arXiv preprint arXiv:2503.10512*.

Huanxin Sheng, Xinyi Liu, Hangfeng He, Jieyu Zhao, and Jian Kang. 2024. Analyzing uncertainty of llm-as-a-judge: Interval evaluations with conformal prediction. *arXiv preprint arXiv:2410.02106*.

Julio Silva-Rodríguez, Ismail Ben Ayed, and Jose Dolz. 2025. Conformal prediction for zero-shot models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jiayuan Su, Jing Luo, Hongwei Wang, and Lu Cheng. 2024. API is enough: Conformal prediction for large language models without logit-access. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1009–1035, Miami, Florida, USA. Association for Computational Linguistics.

Jiashuo Sun, Yi Luo, Yeyun Gong, Chen Lin, Yelong Shen, Jian Guo, and Nan Duan. 2024. Enhancing chain-of-thoughts prompting with iterative bootstrapping in large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, Mexico City, Mexico. Association for Computational Linguistics.

Gauthier Thurin, Kimia Nadjahi, and Claire Boyer. 2025. Optimal transport-based conformal prediction.

Ryan J. Tibshirani, Rina Foygel Barber, Emmanuel J. Candès, and Aaditya Ramdas. 2019. Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems*, volume 32, pages 2526–2536. Curran Associates, Inc.

Alexander Timans, Christoph-Nikolas Straehle, Kaspar Sakmann, and Eric Nalisnick. 2024. Adaptive bounding box uncertainties via two-step conformal prediction. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. 2005. *Algorithmic Learning in a Random World*. Springer, New York, NY.

Nien-Shao Wang, Duygu Nur Yaldiz, Yavuz Faruk Bakman, and Sai Praneeth Karimireddy. 2025a. Conformal prediction adaptive to unknown subpopulation shifts. *Preprint*, arXiv:2506.05583.

Xunguang Wang, Zhenlan Ji, Wenxuan Wang, Zongjie Li, Daoyuan Wu, and Shuai Wang. 2025b. Sok: Evaluating jailbreak guardrails for large language models. *Preprint*, arXiv:2506.10597.

Zhiyuan Wang, Jinhao Duan, Lu Cheng, Yue Zhang, Qingni Wang, Xiaoshuang Shi, Kaidi Xu, Heng Tao Shen, and Xiaofeng Zhu. 2024. Conu: Conformal uncertainty in large language models with correctness coverage guarantees. In *Findings of the Association for Computational Linguistics: EMNLP 2024*.

Zhiyuan Wang, Jinhao Duan, Qingni Wang, Xiaofeng Zhu, Tianlong Chen, Xiaoshuang Shi, and Kaidi Xu. 2025c. COIN: Uncertainty-guarding selective question answering for foundation models with provable risk guarantees. *Preprint*, arXiv:2506.20178.

Zhiyuan Wang, Qingni Wang, Yue Zhang, Tianlong Chen, Xiaofeng Zhu, Xiaoshuang Shi, and Kaidi Xu. 2025d. SConU: Selective conformal uncertainty in large language models. *Preprint*, arXiv:2504.14154. Accepted by ACL 2025 Main.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Gina Wong, Drew Prinster, Suchi Saria, Rama Chellappa, and Anqi Liu. 2025a. Improving coverage in combined prediction sets with weighted p-values. *Preprint*, arXiv:2505.11785.

Gina Wong, Drew Prinster, Suchi Saria, Rama Chellappa, and Anqi Liu. 2025b. Improving coverage in combined prediction sets with weighted p-values. *Preprint*, arXiv:2505.11785.

Jing Xiong, Qiujiang Chen, Fanghua Ye, Zhongwei Wan, Chuanyang Zheng, Chenyang Zhao, Hui Shen, Alexander Hanbo Li, Chaofan Tao, Haochen Tan, Haoli Bai, Lifeng Shang, Lingpeng Kong, and Ngai Wong. 2025. Atts: Asynchronous test-time scaling via conformal prediction. *Preprint*, arXiv:2509.15148.

Rui Xu, Chao Chen, Yue Sun, Parvathinathan Venkitasubramaniam, and Sihong Xie. 2025a. Wasserstein-regularized conformal prediction under general distribution shift. *Preprint*, arXiv:2501.13430.

Yunpeng Xu, Wenge Guo, and Zhi Wei. 2025b. Selective conformal risk control. *Preprint*, arXiv:2512.12844.

Yasin Abbasi Yadkori, Ilja Kuzborskij, David Stutz, András György, Adam Fisch, Arnaud Doucet, Iuliya Beloshapka, Wei-Hung Weng, Yao-Yuan Yang, Csaba Szepesvári, Ali Taylan Cemgil, and Nenad Tomasev. 2024. Mitigating llm hallucinations via conformal abstention. *Preprint*, arXiv:2405.01563.

Margaux Zaffran, Olivier Féron, Yannig Goude, Julie Josse, and Aymeric Dieuleveut. 2022. Adaptive conformal predictions for time series. In *International Conference on Machine Learning*, pages 25834–25866. PMLR.

William Zhang, Saurabh Amin, and Georgia Perakis. 2025. Modular and adaptive conformal prediction for sequential models via residual decomposition. *Preprint*, arXiv:2510.04406.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *Preprint*, arXiv:2307.15043.

6 Appendix

6.1 Discussion

Beyond consolidating recent progress, a key message of this work is that conformal prediction (CP) for LLMs is still missing several "systems-grade" building blocks that could drive the next wave of research. First, semantic nonconformity design remains underdeveloped: RAG-focused CP methods show that end-to-end correctness sets are possible when we calibrate to semantic equivalence rather than exact match (Li et al., 2024b), but the field needs principled score constructions that are robust to judge noise and distribution shift (e.g., judge ensembles, calibration diagnostics, and semantics-aware losses). Second, LLM deployments are multi-stage pipelines (retrieve → generate → verify → act), suggesting a compositional CP theory where error budgets are allocated across stages and guarantees compose; existing retrieval-thresholding approaches (Rouzrokh et al., 2024) provide a template, but extending this to tool use, multi-hop retrieval, and agentic workflows is largely open. Third, multi-source and adaptive aggregation is likely essential in practice (multiple prompts, models, attack families, domains): data-dependent weighting and aggregation can tighten coverage relative to naive unions (Wong et al., 2025b), and developing fast online updates under shift (building on adaptive conformal ideas (Gibbs and Candès, 2021)) is a natural direction for robust LLM safety layers. Fourth, compute-aware conformalization is emerging as a practical necessity: recent work uses CP to control rejection/acceptance in test-time scaling (Xiong et al., 2025) and to produce confidence-controlled early exits during reasoning (Akgül et al., 2025); a broader opportunity is to couple CP with decoding and verification to yield guarantees under strict latency/token budgets. Finally, the community needs generation-native CP primitives: methods that produce valid prediction sets for deep generative models via sampling and admissibility tests (Shahrokhi et al., 2025a) point toward CP that natively handles open-ended outputs (text/code) rather than forcing them into classification-style surrogates. Taken together, these directions suggest that the most impactful advances will come from marrying new CP theory (composition, adaptivity, aggregation) with LLM-specific semantics, pipeline structure, and efficiency constraints.

6.2 Limitation

While this survey provides a systematic synthesis of the burgeoning CP-LLM literature, several constraints regarding its scope and methodology warrant acknowledgement. Chief among these is the qualitative nature of our analysis, which precludes a direct, head-to-head empirical comparison of the reviewed algorithms. Insofar as the field lacks a unified benchmarking platform, the "Pareto efficiency" of different frameworks, specifically the balance between set tightness and inference latency, remains to be rigorously verified through future large-scale experimentation.

Furthermore, the validity of the taxonomies and conclusions presented herein is intrinsically tethered to the quality of surrogate metrics utilized in the primary literature. Despite our discussion of semantic correctness in Section 3, the underlying reliance on auxiliary models for evaluation introduces potential biases that our survey can identify

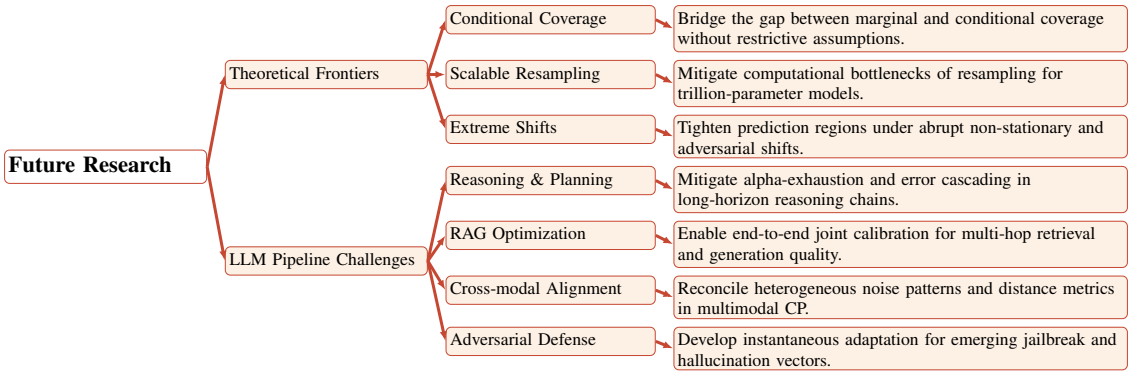

Figure 2: Taxonomy of open challenges and future directions for conformalized LLM systems.

but not empirically rectify. Additionally, the linguistic scope of this work is naturally bounded by the current English-centric landscape of CP research. Notwithstanding our efforts to incorporate diverse applications, the fairness and robustness of conformal guarantees in low-resource and cross-cultural contexts remain under-represented in the existing body of work. Finally, given the unprecedented velocity of LLM development, this survey represents a temporal snapshot; certain emerging non-Transformer architectures may thus require further refinement of the categorization established in our Figure 1.