Journal of Computing and Electronic Information Management
ISSN: 2413-1660 | Vol. 23, No. 2, 2026

Uncertainty Quantification and Abstention for Small Language Models: A Review with an Empirical Case Study on Domain-Specific Question Answering

Tian Lu

Northeastern University, Boston, MA, USA

Abstract: Small language models (SLMs) are increasingly deployed on devices and in domain-specific settings where a wrong answer is costly, so they need to know when not to answer. Existing surveys of uncertainty quantification (UQ) and abstention focus on large general-purpose models. This paper reviews UQ methods and abstention mechanisms from the perspective of SLM deployment and summarizes evidence that small models tend to be overconfident. We complement the review with a case study on medical multiple-choice questions using three 4B-class models. Conformal prediction meets its coverage target when calibration and test data match, but at a 90% target the models abstain on more than half of the questions, and coverage drops when the calibration data come from a different subject or question source. Calibrating separately for each subject corrects the under-coverage of harder subjects but needs several hundred labeled examples per subject to be reliable. We conclude with open problems for reliable abstention in small models.

Keywords: Small language models; Uncertainty quantification; Abstention; Conformal prediction; Selective prediction; Trustworthy AI.

1. Introduction

Language models with fewer than about ten billion parameters, usually called small language models (SLMs), are increasingly the models that are actually deployed. They can run on a single consumer GPU or on a phone, they keep sensitive data on premises, they answer with low latency, and they can be specialized cheaply for a single domain [1, 2]. Recent open-weight families such as Qwen3.5, Gemma 4 and Phi-4 all include models with roughly 2B–4.5B (effective) parameters [3–5].

These deployment settings are frequently also the ones where a wrong answer is costly: clinical decision support, legal or financial assistance, and on-device agents that act without a human in the loop. A small model will be wrong more often than a large one, so the practical question is less "how accurate is it?" than "does it know when it is likely to be wrong, and can the system abstain or defer in those cases?". Several lines of evidence suggest that this question is harder for SLMs. Good calibration of multiple-choice token probabilities has been reported for larger models [6], the post-training that deployed chat models undergo has been reported to degrade calibration [7, 8], with instruction tuning in particular harming it [9], and recent SLM-specific studies report that cheap confidence signals become less reliable as models shrink [10] and that the verbalized confidence of models below about 2B parameters barely separates correct from incorrect answers [11].

Uncertainty quantification (UQ) and abstention for large language models (LLMs) have been surveyed extensively [12–15], including recent surveys focused on agents and multimodal models [16–18]. These surveys treat model size as one factor among many. In this paper we take the SLM deployment setting as the organizing lens and ask which UQ and abstention techniques remain usable, and reliable, under its constraints: limited compute for sampling, white-box access to logits, frequent quantization or distillation, and narrow, domain-specific use.

Our contributions are threefold:

1. **A taxonomy** of UQ methods (Section 3) and abstention/deferral mechanisms (Section 4), annotated with the access each method needs, its inference cost, and the issues reported specifically for small models (Table 2).
2. **A synthesis of SLM-specific evidence** on calibration, the effect of post-training, quantization and distillation, and SLM-to-LLM routing, together with the evaluation metrics and benchmarks used in this literature (Sections 3–5).
3. **A reproducible empirical case study** (Section 6) of conformal abstention for three 4B-class SLMs from different families on medical multiple-choice questions. We measure calibration, test whether conformal prediction delivers its coverage guarantee in distribution, quantify how coverage fails when the calibration data come from a different medical subject or from a different exam source, and evaluate subject-conditional (Mondrian) calibration as a fix.

2. Background and Problem Formulation

2.1. Uncertainty, calibration and abstention

Aleatoric uncertainty is the irreducible ambiguity of the task itself (e.g., a question with two defensible answers), whereas *epistemic* uncertainty stems from the model's lack of knowledge and could in principle be reduced with more data or a better model. For abstention, epistemic uncertainty is usually the quantity of interest: a model should decline questions it does not know, not merely ones that are ambiguous.

A confidence score is *calibrated* if, among all predictions made with confidence c, a fraction c is correct.

Calibration is a property of the score's absolute values; *discrimination* (whether correct predictions receive higher confidence than incorrect ones) is a property of its ranking. A score can rank well and still be badly calibrated, which is why post-hoc recalibration (Section 3.5) is often sufficient to fix the former but not the latter.

Selective prediction (also called classification with a reject option) lets a model abstain on some inputs in exchange for higher accuracy on the rest [19–21]. *Deferral* is the variant in which abstained inputs are passed to another decision-maker, such as a larger model or a human expert [22].

2.2. Notation

Let x be an input (a question with its answer options), $\mathcal{Y}$ the set of candidate answers, and $\hat{p}(y \mid x)$ the model's probability for answer $y \in \mathcal{Y}$. The prediction is $\hat{y}(x) = \text{argmax}_y \hat{p}(y \mid x)$, and a confidence function $c(x)$ maps the input to a scalar, for instance $c(x) = \max_y \hat{p}(y \mid x)$. An abstention function $r(x) \in \{0,1\}$ decides whether to answer $(r = 1)$ or abstain $(r = 0)$; threshold rules use $r(x) = \mathbb{1}[c(x) \geq \tau]$.

2.3. Selective risk, coverage and conformal guarantees

For an abstention rule r, *coverage* is the fraction of inputs answered and *selective risk* is the error rate on those inputs:

$$\phi(r) = \mathbb{E}[r(x)], \qquad R(r) = \frac{\mathbb{E}[r(x)\,\mathbb{1}[\hat{y}(x) \neq y]]}{\phi(r)}.$$

Sweeping τ traces the risk–coverage curve [20]; the area under it (AURC) summarizes how well $c(x)$ ranks correct above incorrect answers.

Conformal prediction [23, 24] returns a *set* $C(x) \subseteq \mathcal{Y}$ instead of a single answer. In split conformal prediction [25], a score $s(x, y)$ is computed on n held-out calibration examples. With the least-ambiguous set-valued classifier (LAC) score [26], $s_i = 1 - \hat{p}(y_i \mid x_i)$, the threshold $\hat{q}$ is the $\lceil (n+1)(1-\alpha) \rceil$ -th smallest calibration score, and the prediction set is

$$C(x) = \{y : \hat{p}(y \mid x) \geq 1 - \hat{q}\}.$$

If calibration and test examples are exchangeable, this guarantees marginal coverage

$$\Pr[y_{\text{test}} \in C(x_{\text{test}})] \geq 1 - \alpha$$

for any model, however poorly calibrated. The guarantee is marginal (averaged over inputs) and holds only under exchangeability; both caveats are central to our case study. A natural abstention rule answers only when $C(x)$ contains exactly one answer.

2.4. Related surveys

Table 1 positions this review relative to existing surveys. Shorinwa et al. [12] and Liu et al. [13] provide general taxonomies of LLM UQ, the latter organized by computational cost and by the dimension of uncertainty (input, reasoning, parameter, prediction). Wen et al. [14] survey abstention for LLMs from the perspectives of the query, the model and human values. Zhang et al. [15] trace how uncertainty is being used as an active control signal for reasoning, agents and reinforcement learning. Oh et al. [16] and Lin et al. [17] focus on LLM agents, and Boudiaf et al. [18] on multimodal LLMs. Campos et al. [27] survey conformal prediction for NLP. None of these takes the small-model setting as its focus, and none reports new experiments on SLMs. Conversely, the SLM surveys of Wang et al. [1] and Lu et al. [2] cover SLM design, training and efficiency (and, in the case of Wang et al., trustworthiness more broadly), but treat UQ and abstention only briefly. Our review sits between these two bodies of work.

Table 1. Existing surveys and how this review differ

Survey	Scope	Difference from this review
Shorinwa et al. [12]	General LLM UQ taxonomy and applications	Not SLM-focused; no experiments
Liu et al. [13]	LLM UQ by cost and uncertainty dimension	Not SLM-focused; no experiments
Wen et al. [14]	LLM abstention (query / model / human values)	No SLM-specific treatment of calibration or deployment constraints
Zhang et al. [15]	Uncertainty as a control signal	Not SLM-focused
Oh et al. [16]; Lin et al. [17]	UQ for LLM agents	Agent- and large-model-centred
Boudiaf et al. [18]	Uncertainty in multimodal LLMs	Different modality
Campos et al. [27]	Conformal prediction in NLP	Not LLM- or SLM-specific; no abstention study
Wang et al. [1]; Lu et al. [2]	Small language models (incl. trustworthiness)	UQ and abstention covered only briefly

3. Uncertainty Quantification Methods for SLMs

Figure 1 summarizes the taxonomy used in Sections 3 and 4, and Table 2 compares the method families along the dimensions that matter for SLM deployment.

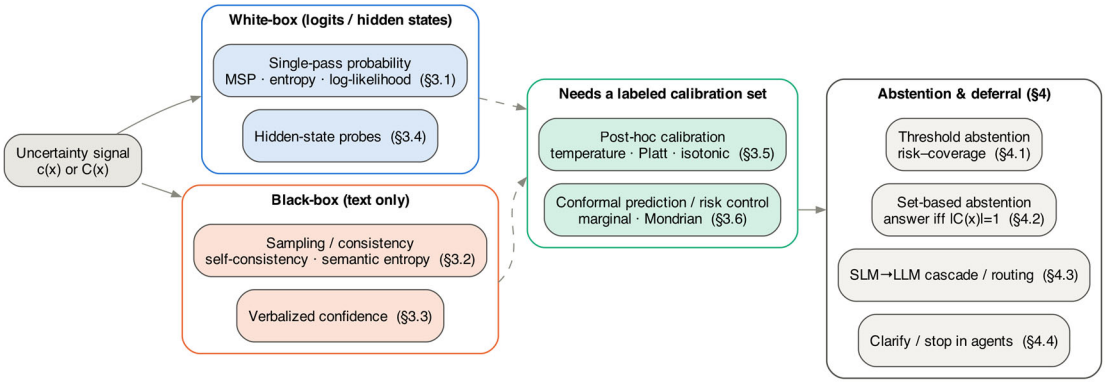


Figure 1. Taxonomy of uncertainty quantification and abstention methods discussed in this review. Families are grouped by the access they need to the model; the right column lists the abstention and deferral mechanisms that consume the uncertainty signal.

3.1. Single-pass probability-based methods

The cheapest signals are read from one forward pass: the maximum softmax probability (MSP) of the predicted answer, the entropy of the answer distribution, or, for free-form generation, the (length-normalized) sequence log-likelihood. MSP was proposed as a baseline for detecting misclassified and out-of-distribution inputs [28] and remains a strong default. For multiple-choice questions, Kadavath et al. [6] reported that larger LMs are well calibrated when the question is presented in an appropriate format, and the GPT-4 report found that the pre-trained model's answer-letter log-probabilities were highly calibrated on an MMLU subset but became markedly less so after post-training [7]. Earlier, Jiang et al. [29] found the QA probabilities of T5, BART and GPT-2 to be poorly calibrated.

For SLMs, these signals have two attractive properties: they require no extra computation, and SLMs are usually run locally with full white-box access to logits. Their weakness is reliability at small scale: a cross-scale study of eight 0.5B–7B models (available only as a preprint) reports that logit-based confidence saturates and becomes overconfident and that normalized entropy discriminates failures barely better than chance at 1B parameters and below [10].

3.2. Sampling and consistency methods

Sampling-based methods draw several answers and measure their agreement. Self-consistency takes a majority vote over sampled reasoning paths [30], and the vote share can serve as a confidence score. Semantic entropy clusters sampled answers by meaning before computing entropy, so that paraphrases of the same answer are not counted as disagreement [31]; it has been shown to detect confabulations across datasets without task-specific training [32]. In the SLM setting, consistency signals appear to degrade more gracefully with scale than logit-based ones [10], and answer stability under a perturbed prompt (adding an "I don't know" option) has been proposed as an abstention signal for 2B–8B models [33]. The cost is k forward passes per query, which partly cancels the latency and energy advantages that motivate SLM deployment in the first place.

3.3. Verbalized confidence

Verbalized methods ask the model to state its confidence in words or numbers. Lin et al. [34] showed that GPT-3 can be fine-tuned to express calibrated verbal probabilities, and for large RLHF-tuned API models, verbalized confidence can be better calibrated than token probabilities [35]. However, a broad benchmark of black-box elicitation strategies found LLMs to be overconfident when verbalizing confidence, and all elicitation methods struggled on tasks requiring professional knowledge [36], and agent studies found verbalized confidence unreliable as an uncertainty proxy [37]. Evidence specific to small models is less favourable still: across eleven instruction-tuned models from 0.5B to 14B, Shen [11] reports cases where stated confidence stays above 0.5 while accuracy is below it, a pattern that temperature scaling cannot repair, and that even after Platt scaling only a few model–task pairs could be certified for autonomous operation at a 20% risk level. Verbalized confidence is attractive because it works through a text-only interface, but for SLMs it should not be the default signal.

3.4. Internal-representation probes

When hidden states are accessible, a light classifier can be trained on them to predict whether an answer or statement is correct. Azaria and Mitchell [38] trained such probes to predict the truthfulness of statements with 71–83% accuracy depending on the base model, and Kadavath et al. [6] trained models to predict whether they know the answer, observing that this ability generalizes imperfectly to new tasks. Kapoor et al. [39] argue that prompting alone is insufficient and that fine-tuning on roughly a thousand graded examples yields uncertainty estimates that generalize. Probes suit SLMs well, because training them is cheap and SLMs are typically self-hosted, but they need labeled in-domain data, which is precisely what is scarce in many specialized deployments (Section 7).

3.5. Post-hoc calibration and the effect of model compression

Post-hoc calibration maps raw scores to calibrated probabilities using a held-out labeled set. Platt scaling fits a logistic function to the scores [40]; temperature scaling, its single-parameter multiclass variant, rescales logits by a scalar and was found to be surprisingly effective for modern

neural networks [41]; Zadrozny and Elkan [42] proposed calibrating any ranking classifier and combining binary estimates into multiclass probabilities; and Naeini et al. [43] proposed Bayesian Binning into Quantiles and evaluated calibration with the expected and maximum calibration errors (ECE, MCE). Because these maps are monotone, they change calibration but not the ranking of answers, so they cannot fix a confidence score that discriminates poorly.

The training pipeline of a deployed SLM matters here. Post-training can reduce calibration [7]: Zhu et al. [9] study how calibration evolves through pretraining and alignment, Leng et al. [8] attribute verbalized overconfidence after RLHF to biased reward models, and Huang et al. [44] show that instruction tuning on high-resource data raises confidence in low-resource languages without matching accuracy gains. Steps that are common in SLM pipelines also affect confidence: 4-bit quantization, typical for edge deployment, lowers confidence on the true label, especially on examples that were already uncertain [45], and on-policy distillation used in post-training can induce severe overconfidence [46]. Many small models are also trained with signals from larger ones; for example, Llama 3.2's 1B and 3B models used larger Llama 3.1 models' logits during pretraining [47]. Calibration measured on the full-precision checkpoint therefore need not hold for the quantized model that is actually shipped.

3.6. Distribution-free methods: conformal prediction and risk control

Conformal prediction wraps any model, regardless of its calibration, and produces prediction sets with a finite-sample coverage guarantee under exchangeability [23–25, 48]. Different conformity scores trade off set size and adaptivity: LAC yields the smallest sets on average [26]; APS targets better approximate conditional coverage [49], and RAPS regularizes it to obtain smaller, more stable sets [50]. Conformal risk control generalizes the guarantee from miscoverage to any monotone loss [51], and conformal language modeling extends it to open-ended generation by calibrating sampling and rejection rules [52]. For multiple-choice QA, Kumar et al. [53] found that conformal uncertainty estimates correlate tightly with accuracy and examined whether exchangeability holds for questions from other subjects; Ye et al. [54] used conformal prediction to benchmark uncertainty across nine LLM series and found that larger models are not necessarily more certain.

Two limitations are directly relevant to deployment. First, the guarantee is marginal: coverage can be far below $1-\alpha$ for specific subgroups even when the average is correct. Mondrian (group-conditional) conformal prediction calibrates separately within each group and restores coverage per group, at the price of needing calibration data for every group [55]. Second, exchangeability fails under distribution shift. Weighted conformal prediction handles covariate shift when the likelihood ratio can be estimated [56], and more general bounds quantify the coverage loss under arbitrary drift [57]. For SLMs, conformal methods are attractive because they need only a single forward pass and a few hundred labeled examples, and because they turn even a miscalibrated score into a guarantee. Our case study (Section 6) tests how robust that guarantee is in practice.

Table 2. UQ method families from the perspective of SLM deployment.

Family	White-box access	Inference cost	Known issues on SLMs	Representative references
Single-pass probability (MSP, entropy, log-likelihood)	Logits	1 pass	Overconfident; entropy discriminates poorly at ≤1B	[6, 10, 28]
Sampling / consistency (self-consistency, semantic entropy)	None (black-box)	k passes (+ clustering)	Robust across scale but cost erodes SLM efficiency	[30–33]
Verbalized confidence	None	1 pass (longer output)	Overconfident; in some small models confidence > 0.5 with accuracy < 0.5, not fixable by temperature scaling	[11, 34, 36]
Hidden-state probes	Hidden states + labels	1 pass + probe	Needs labeled in-domain data	[38, 39]
Post-hoc calibration (temperature, Platt, isotonic)	Logits + labels	Negligible	Fixes calibration, not ranking; may break after quantization	[40–42]
Conformal prediction / risk control	Scores + labels	1 pass	Marginal guarantee only; fails without exchangeability	[24, 26, 51, 55]

4. Abstention and Deferral Mechanisms

4.1. Threshold-based abstention and the risk–coverage trade-off

The simplest abstention rule answers only when $c(x) \geq \tau$. Chow [19] analysed the optimal error–reject trade-off, El-Yaniv and Wiener [20] formalized the risk–coverage trade-off, and Geifman and El-Yaniv [21] showed how to pick a threshold for a deep network that meets a user-specified risk with high probability. For LLMs, abstaining with an appropriate uncertainty measure has been reported to improve correctness and to avoid a large share of hallucinations on unanswerable questions [58]; self-evaluation learned through parameter-efficient tuning improves selective prediction [59]; and calibration- or prompting-based abstention baselines can be beaten by having several LLMs probe each other's answers [60]. For SLMs, a threshold chosen on a calibration set is only as good as the ranking quality of $c(x)$ (Section 3). A plug-in threshold also carries no finite-sample guarantee, and thresholds reused on a different task can violate the risk budget [11].

4.2. Abstention from conformal prediction sets

Conformal sets give a natural abstention rule: answer

when the set is a singleton and abstain (or show all candidates) otherwise. Unlike a raw threshold, the threshold $1 - \hat{q}$ is chosen so that the *set* covers the truth with probability at least $1 - \alpha$. Note that this does not directly bound the error among answered questions; the accuracy on singletons is an empirical quantity that must be reported alongside coverage. Conformal sets have also been used to decide when a language-model planner should ask a human for help [61], connecting set size to clarification (Section 4.4).

4.3. SLM-to-LLM cascades and routing

Instead of abstaining outright, an SLM can defer uncertain queries to a larger model, which is the natural deployment pattern for on-device assistants. Learning to defer [22] provides the decision-theoretic foundation. FrugalGPT selects among LLM APIs of different cost per query [62]; Hybrid LLM routes queries between a small and a large model based on predicted difficulty [63]; and RouteLLM trains routers on human preference data [64]. Gupta et al. [65] show that, for small-to-large cascades, sequence-level uncertainty is biased by output length and that learned deferral rules over token-level uncertainties work better. Most directly relevant, Chuang et al. [66] benchmark uncertainty-based routing from on-device SLMs to LLMs and find that routing quality depends on how well uncertainty aligns with correctness, and that uncertainty distributions depend more on the SLM and the UQ method than on the downstream data. This last observation suggests that a routing threshold tuned for one SLM may not transfer to another.

4.4. Clarification and stopping in agentic settings

In agents, abstention becomes a sequential decision: whether to call a tool, ask the user for clarification, or stop. UALA uses uncertainty to decide when a language agent should consult external tools, improving over ReAct with fewer tool calls [37]. KnowNo uses conformal prediction to let a robot planner ask for help with a statistical guarantee on task success [61]. Davidov et al. [67] frame mid-generation abstention as an action in reinforcement learning, and Luo et al. [68] find that LLM agents often abstain too late or not at all, with larger models sometimes abstaining less promptly. Agent-oriented surveys identify uncertainty propagation across steps as a key open problem [16, 17]. SLM-based agents are attractive for on-device automation, but how per-step uncertainty of a small model should be aggregated over a trajectory is largely unexplored.

5. Evaluation Metrics and Benchmarks

Calibration metrics. The expected calibration error (ECE) bins predictions by confidence and averages the gap between confidence and accuracy across bins [41, 43]; it depends on the binning scheme, which should therefore always be reported. The Brier score [69] is the squared error between the predicted distribution and the one-hot label; unlike ECE, it does not depend on a choice of bins.

Discrimination and selective prediction metrics. AUROC measures how well confidence separates correct from incorrect answers, independently of calibration. The risk–coverage curve [20, 21] and the area under it (AURC) summarize the full abstention trade-off. For conformal methods, the relevant quantities are empirical coverage, average set size, and the fraction of singleton sets (the answer rate under singleton abstention), together with accuracy on answered questions.

Benchmarks and tools. Multiple-choice benchmarks are convenient for UQ because answer probabilities can be read directly from the option letters: MMLU [70], ARC [71], and, for medicine, MedQA [72] and MedMCQA [73]. TruthfulQA targets imitative falsehoods [74]. LM-Polygraph [75] and the benchmark of Vashurin et al. [76] implement many UQ methods under a common interface, and Ye et al. [54] benchmark LLMs with conformal prediction. Few of these benchmarks report results stratified by subject or by data source, which is exactly where marginal guarantees can hide failures.

6. Empirical Case Study: Conformal Abstention under Subject and Source Shift

The review identifies conformal prediction as the method family whose assumptions are easiest to state and, for SLMs, cheapest to satisfy: one forward pass and a few hundred labeled examples. Its guarantee, however, is marginal and requires exchangeability (Section 3.6), and deployments of domain-specific SLMs rarely have calibration data that match the test distribution exactly. We therefore ask four questions. **RQ1:** How well calibrated are current 4B-class SLMs on medical multiple-choice questions? **RQ2:** Does conformal prediction reach its target coverage in distribution, and what does abstention cost? **RQ3:** How much coverage is lost when calibration and test questions come from different medical subjects? **RQ4:** Does subject-conditional (Mondrian) calibration restore coverage, and how much calibration data does it need? We additionally test a natural shift in question source.

6.1. Setup

Data. We use MedMCQA [73], whose released data provide, for each four-option question, a single correct-option field and a subject label. Its splits are defined by exam source: the training split consists of mock and online test series, whereas the development (validation) split consists of questions from the NEET-PG entrance exam [73]. Because the public test split has no labels, and because calibrating on one source and testing on another would violate exchangeability by construction, we build an **in-distribution pool** by sampling 2,000 questions per subject from the training split (no model is trained on it), and use the 2,132 validation questions of the same subjects as a **real-exam set** for the source-shift experiment. We study four subjects with at least 200 validation questions each: Anatomy, Dental, Pathology and Pharmacology. Questions were deduplicated by normalized text, and training questions whose stem also appears in the validation split were removed. Although MedMCQA's `choice_type` field marks some questions as multi-choice, every question in the released data has exactly one correct option, so no question was excluded on that basis.

Models. We evaluate one recent instruction-tuned model from each of three families at a similar scale: Qwen3.5-4B [3], Gemma-4-E4B-it [4] and Phi-4-mini-instruct (3.8B) [5]. A Llama model was planned but not included because access to the gated Llama-3.2-3B-Instruct weights was not obtained. Models were run in bfloat16 with Hugging Face Transformers 5.22.2 on one NVIDIA RTX A5000 GPU; exact model revisions were recorded. Thinking modes were disabled through each model's chat template.

Confidence extraction. Each question is formatted with its four options labeled A–D, followed by the instruction "Respond with only the letter (A, B, C, or D) of the correct option. Do not explain.", and wrapped in the model's chat template. From a single forward pass we read the next-token distribution at the start of the assistant turn and sum, for each letter, the probabilities of its tokenizations with and without a leading space; renormalizing over the four letters gives $\hat{p}(y \mid x)$. The instruction was fixed after a 100-question pilot on the criterion that the four letters receive more than 90% of the next-token probability (the instruction originally proposed failed this for Gemma 4, which often began with "The answer is"). On the full data, the four letters received on average at least 99.8% of the probability mass for every model. No sampling is involved anywhere, so the outputs are deterministic.

Methods. B1 abstains when the maximum softmax probability is below a threshold chosen on the calibration set as the smallest threshold whose empirical selective risk is at most α; B2 does the same after temperature scaling fitted on the calibration set [41]. M1 is split conformal prediction with the LAC score [26] calibrated on a pooled calibration set; M2 is the Mondrian variant [55], which computes a separate threshold $\hat{q}$ for each subject. M1 and M2 answer only when the prediction set is a singleton.

Experiments. E1 evaluates calibration on the full pools. E2 draws $n = 1000$ calibration questions from the in-distribution pool and tests on the remaining 7,000, for $\alpha \in \{0.05, 0.1, 0.2\}$. E3 calibrates on 500 questions of subject i and tests on subject j, giving a 4×4 coverage matrix. E4 compares M1 calibrated on $4n$ subject-balanced questions with M2 calibrated on n questions per subject, for $n \in \{50, 100, 200, 500\}$. E5 calibrates M1 and M2 on the in-distribution pool (500 per subject) and tests on the real-exam set. Every setting is repeated over 20 random calibration draws (seeds 0–19) and reported as mean ± standard deviation. All tables and figures are computed by a single script from the saved per-question probabilities.

6.2. RQ1: Calibration

Table 3. Calibration and discrimination of the three models.

Model	Pool	n	Acc.	Mean conf.	ECE	Brier	AUROC	AURC
Qwen3.5-4B	Train pool	8000	0.665	0.748	0.083	0.443	0.809	0.138
Qwen3.5-4B	Validation (real exams)	2132	0.568	0.686	0.120	0.571	0.739	0.240
Gemma-4-E4B-it	Train pool	8000	0.618	0.861	0.243	0.599	0.771	0.184
Gemma-4-E4B-it	Validation (real exams)	2132	0.534	0.832	0.298	0.713	0.733	0.269
Phi-4-mini-instruct	Train pool	8000	0.584	0.702	0.119	0.549	0.757	0.218
Phi-4-mini-instruct	Validation (real exams)	2132	0.470	0.656	0.186	0.669	0.706	0.354

Note. ECE uses 15 equal-width bins; the Brier score is summed over the four options.

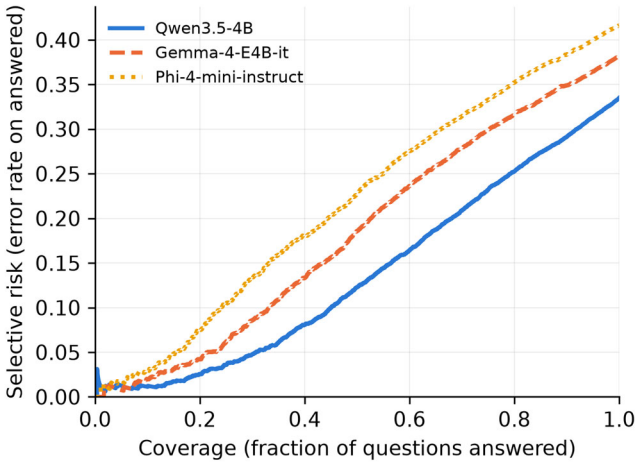


Figure 2. Risk–coverage curves on the in-distribution pool (8,000 questions), using the maximum option probability as confidence. Lower is better.

All three models are overconfident: on the in-distribution pool, mean confidence exceeds accuracy by 0.08 (Qwen3.5-4B), 0.24 (Gemma-4-E4B-it) and 0.12 (Phi-4-mini-instruct) (Table 3). Calibration and discrimination do not rank the models in the same order: Gemma-4-E4B-it has by far the largest ECE (0.243) but a higher AUROC than Phi-4-mini-instruct (0.771 vs. 0.757), which illustrates why a poorly calibrated score can still be useful for abstention once it is recalibrated or conformalized. Accuracy is consistently higher on Pharmacology and Pathology than on Anatomy and Dental for all three models. On the real-exam questions, accuracy falls by 8.4–11.4 points and ECE increases for every model, so a confidence threshold tuned on one question source would answer more

questions wrongly on the other.

Figure 2 shows the risk–coverage curves. Qwen3.5-4B dominates over the whole range; at full coverage its error rate is 0.335 versus 0.382 and 0.416 for the other two models. Temperature scaling (B2) changed the answer rate of threshold-based abstention by at most 0.006 relative to B1 in any setting, as expected for a transformation that leaves the ranking of questions almost unchanged.

6.3. RQ2: In-distribution conformal prediction

Conformal prediction achieves its nominal coverage for every model and every α, with mean coverage within 0.007 of $1-\alpha$ (Table 4). This holds for Gemma-4-E4B-it as well, despite its ECE of 0.243: the guarantee does not depend on calibration. The price is a low answer rate. At a 90% coverage target, only 44.7%, 37.4% and 29.2% of questions receive a singleton set, and at $\alpha = 0.05$ average set sizes exceed 2.4 options out of 4. A second observation concerns the abstention rule itself: accuracy on the answered (singleton) questions is not controlled at $1-\alpha$. At $\alpha = 0.05$ it is 0.911–0.937, below the 0.95 that a user might expect. The threshold baseline B1, which directly targets selective risk, reaches about 0.95 accuracy on answered questions on average, but its answer rate varies two to four times more across calibration draws than that of M1 at $\alpha = 0.05$, and its plug-in threshold carries no finite-sample guarantee. Practitioners who need a bound on the error of answered questions should therefore use procedures that certify selective risk [21] or related risk-control methods [51] rather than equating singleton conformal sets with guaranteed answers.

Table 4. In-distribution conformal and threshold results.

Model	α	Coverage (M1)	Set size (M1)	Answer rate (M1)	Acc. answered (M1)	Answer rate (B1)	Acc. answered (B1)
Qwen3.5-4B	0.05	0.950 ± 0.006	2.44 ± 0.06	0.341 ± 0.013	0.937 ± 0.004	0.308 ± 0.046	0.949 ± 0.014
Qwen3.5-4B	0.1	0.901 ± 0.010	1.99 ± 0.06	0.447 ± 0.016	0.901 ± 0.009	0.445 ± 0.020	0.903 ± 0.011
Qwen3.5-4B	0.2	0.804 ± 0.012	1.48 ± 0.05	0.623 ± 0.022	0.817 ± 0.011	0.668 ± 0.025	0.806 ± 0.012
Gemma-4-E4B-it	0.05	0.951 ± 0.008	2.65 ± 0.08	0.276 ± 0.016	0.924 ± 0.009	0.218 ± 0.034	0.951 ± 0.012
Gemma-4-E4B-it	0.1	0.903 ± 0.011	2.23 ± 0.07	0.374 ± 0.018	0.878 ± 0.012	0.333 ± 0.026	0.899 ± 0.015
Gemma-4-E4B-it	0.2	0.807 ± 0.010	1.71 ± 0.04	0.535 ± 0.017	0.794 ± 0.011	0.517 ± 0.030	0.807 ± 0.018
Phi-4-mini-instruct	0.05	0.950 ± 0.006	2.81 ± 0.05	0.220 ± 0.011	0.911 ± 0.008	0.155 ± 0.032	0.950 ± 0.015
Phi-4-mini-instruct	0.1	0.903 ± 0.010	2.42 ± 0.06	0.292 ± 0.011	0.872 ± 0.007	0.241 ± 0.026	0.901 ± 0.017
Phi-4-mini-instruct	0.2	0.801 ± 0.015	1.82 ± 0.06	0.445 ± 0.020	0.792 ± 0.013	0.443 ± 0.032	0.800 ± 0.017

Note. Calibration $n = 1000$, 20 draws. Answer rate is the fraction of singleton sets (M1) or of questions above the threshold (B1).

6.4. RQ3: Coverage under subject shift

Figure 3 shows the gap between empirical coverage and the 0.90 target when calibration and test subjects differ. On the diagonal (same subject) the gap is within ±0.005 for all models, as the theory predicts. Off the diagonal, coverage depends systematically on relative difficulty: calibrating on an easier subject (Pharmacology or Pathology) and testing on a harder one (Dental or Anatomy) under-covers, while the reverse over-covers and produces larger sets than necessary. The worst case is Qwen3.5-4B calibrated on Pharmacology and tested on Dental, with coverage 0.075 below the target; the most negative gaps for Gemma-4-E4B-it and Phi-4-mini-instruct are 0.029 and 0.048. Shift hurts more at looser targets: at $\alpha = 0.2$ the worst gaps grow to 0.176, 0.079 and 0.129 respectively. Subject shift therefore turns the conformal guarantee into an optimistic estimate exactly when an easier calibration domain is used for a harder deployment domain.

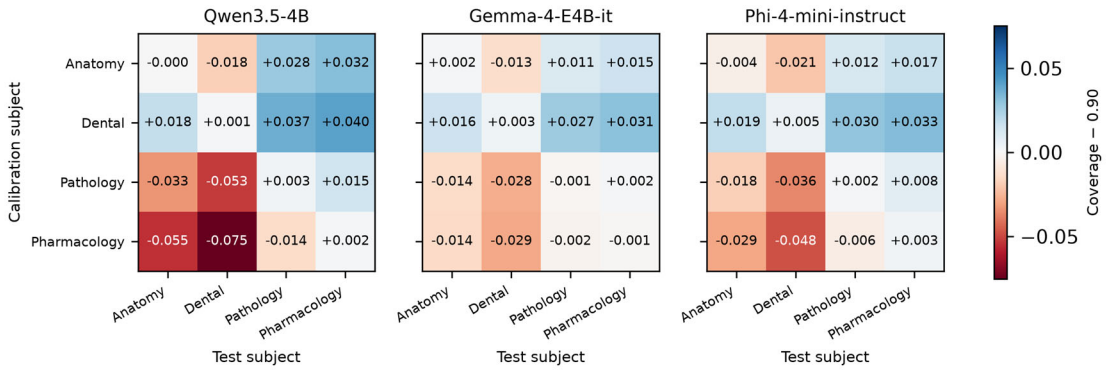


Figure 3. Coverage gap (empirical coverage minus 0.90) when calibrating on one subject (rows, $n = 500$) and testing on another (columns), averaged over 20 draws. Red cells under-cover.

6.5. RQ4: Subject-conditional (Mondrian) calibration

Even with a subject-balanced calibration set, pooled calibration (M1) distributes coverage unevenly across subjects: in every model and at every n, Dental is under-covered (mean coverage 0.863–0.885) and Pharmacology over-covered (0.907–0.931), while marginal coverage stays at 0.90. Mondrian calibration (M2) removes this systematic bias: its mean Dental coverage is 0.892–0.907 at every n

(Table 5), and the answer rate is essentially unchanged. However, each M2 threshold is estimated from only n examples, so its variance across draws is larger. At $n = 50$, the worst subject in a given draw is covered at only 0.847–0.858 under M2, lower than under M1 (0.867–0.881). Only at $n = 500$ does M2 also improve this per-draw worst case for all three models (0.885–0.888 vs. 0.872–0.884; Figure 4). In short, Mondrian calibration corrects the bias of pooled calibration at any sample size, but a few hundred labeled examples per group are needed before it also gives a more reliable worst-group coverage in a single deployment.

Table 5. Pooled vs. Mondrian calibration.

n per subject	Method	Qwen3.5 Dental	Qwen3.5 worst	Gemma Dental	Gemma worst	Phi Dental	Phi worst
50	M1	0.867 ± 0.023	0.867 ± 0.023	0.881 ± 0.018	0.881 ± 0.018	0.873 ± 0.035	0.873 ± 0.035
50	M2	0.901 ± 0.041	0.847 ± 0.025	0.905 ± 0.028	0.853 ± 0.039	0.892 ± 0.056	0.858 ± 0.043
100	M1	0.878 ± 0.017	0.878 ± 0.017	0.885 ± 0.016	0.885 ± 0.016	0.880 ± 0.025	0.880 ± 0.025
100	M2	0.907 ± 0.022	0.875 ± 0.023	0.896 ± 0.027	0.866 ± 0.027	0.900 ± 0.027	0.873 ± 0.022
200	M1	0.863 ± 0.014	0.863 ± 0.014	0.878 ± 0.012	0.878 ± 0.012	0.876 ± 0.013	0.876 ± 0.013
200	M2	0.892 ± 0.028	0.870 ± 0.019	0.892 ± 0.023	0.870 ± 0.017	0.893 ± 0.029	0.877 ± 0.021
500	M1	0.872 ± 0.010	0.872 ± 0.010	0.884 ± 0.009	0.884 ± 0.009	0.880 ± 0.012	0.880 ± 0.012
500	M2	0.901 ± 0.017	0.885 ± 0.010	0.903 ± 0.012	0.888 ± 0.011	0.905 ± 0.014	0.887 ± 0.013

Note. Pooled (M1, $4n$ subject-balanced calibration questions) vs. Mondrian (M2, n per subject) at $\alpha = 0.1$. *Dental*: mean coverage on the hardest subject. *Worst*: per-draw minimum of the four per-subject coverages.

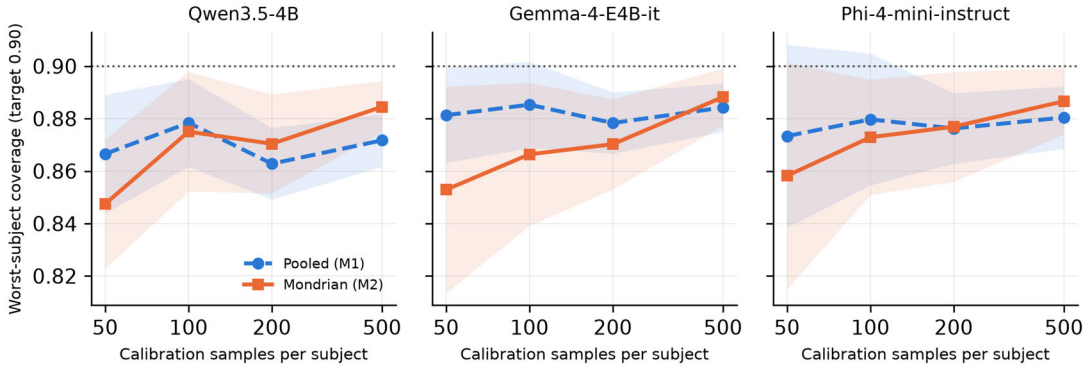


Figure 4. Worst-subject coverage per calibration draw (mean ± s.d. over 20 draws) for pooled (M1) and Mondrian (M2) calibration at $\alpha = 0.1$. The dotted line is the 0.90 target.

6.6. Shift in question source

When the conformal threshold is calibrated on mock-exam questions and applied to real-exam questions of the same subjects, marginal coverage falls to 0.869–0.883, 1.7–3.1 points below the target (Table 6). Part of this shortfall is due to a change in subject mix: Dental, the hardest subject, makes up 62% of the real-exam set but only 25% of the calibration data. Mondrian calibration corrects this component (Dental coverage 0.894–0.900) and raises marginal coverage to 0.884–0.896, but it does not remove the effect of the change in source within a subject; for example, Qwen3.5-4B's coverage on Pharmacology remains at 0.849. Both kinds of shift, in subject and in source, occur naturally when an SLM is calibrated on whatever labeled data are available and then deployed on a specific clinical question stream.

Table 6. Coverage under question-source shift.

Model	Method	Marginal coverage	Dental coverage	Answer rate
Qwen3.5-4B	M1	0.869 ± 0.007	0.860 ± 0.009	0.334 ± 0.009
Qwen3.5-4B	M2	0.884 ± 0.009	0.894 ± 0.014	0.321 ± 0.013
Gemma-4-E4B-it	M1	0.883 ± 0.008	0.874 ± 0.010	0.289 ± 0.013
Gemma-4-E4B-it	M2	0.896 ± 0.007	0.898 ± 0.011	0.274 ± 0.011
Phi-4-mini-instruct	M1	0.882 ± 0.008	0.875 ± 0.009	0.214 ± 0.009
Phi-4-mini-instruct	M2	0.894 ± 0.008	0.900 ± 0.012	0.210 ± 0.010

Note. Calibration on the in-distribution pool (500 per subject), test on the real-exam set, $\alpha = 0.1$.

6.7. Limitations

The case study uses one model of roughly 4B (effective) parameters from each of three families, so it supports comparisons across families at one scale but no conclusions about how uncertainty behaves as model size changes. It covers multiple-choice questions with white-box access to option probabilities, a single dataset and four of its subjects, one frozen prompt, and the LAC score only; adaptive scores such as APS [49] or RAPS [50] may distribute coverage differently. Our in-distribution pool comes from MedMCQA's training split; we cannot rule out that some questions appeared in the models' pretraining data, which could inflate accuracy, although this does not affect the

validity of the conformal guarantee. Finally, the source-shift experiment confounds the change of exam source with any other differences between the splits.

7. Open Challenges and Future Directions

Coverage under distribution shift. Our case study shows that conformal guarantees degrade in exactly the situation that is typical for specialized SLMs: calibration data from one subject or source, deployment on another. Weighted conformal prediction [56] and bounds for non-exchangeable data [57] address this in principle, but they require estimating a likelihood ratio between calibration and deployment inputs or bounding the shift, both of which are hard for text. Practical, label-efficient procedures for detecting and correcting shift in SLM deployments remain an open problem.

Scarcity of domain-specific calibration data. Conditional guarantees need labeled data per group. In our experiments, Mondrian calibration needed several hundred examples per subject before it reliably improved worst-group coverage. Specialized domains such as medicine have many fine-grained groups (specialties, institutions, languages) and few labels, so methods that share statistical strength across groups, or that combine a small amount of target data with larger out-of-domain calibration sets, are needed.

Choosing the signal for small models. Evidence that verbalized confidence is unreliable for small models [11], in line with the general overconfidence of verbalized confidence [36], and that cheap logit-based signals degrade at the smallest scales [10] suggests that SLMs need different defaults than API-served LLMs. Since SLMs are usually self-hosted, white-box signals (option probabilities, hidden-state probes [38]) are available and inexpensive, but apart from a preprint comparing logit-based and sampling signals across scales [10] and a routing-oriented benchmark that includes trained hidden-state probes for 1B–8B models [66], we are not aware of a systematic comparison of these signals together with calibration and conformal guarantees in a specific deployment domain.

Calibrating the model that is actually shipped. Quantization, standard for edge deployment, changes confidence [45], and so can distillation-based post-training [46]. Calibration sets, thresholds and conformal quantiles should be computed for the deployed checkpoint, and it is open how much of a calibration computed at full precision transfers to compressed variants.

Routing and multi-step agents. Uncertainty distributions in SLM-to-LLM routing depend mainly on the SLM and the UQ method [66], which suggests that routing thresholds may not transfer across models. In agents, uncertainty must be propagated over many steps and tool calls [16, 17], and current agents often fail to stop in time [68]. Extending set-valued guarantees such as those of KnowNo [61] to long trajectories executed by small models is a promising direction.

8. Conclusion

We reviewed uncertainty quantification and abstention from the perspective of small language models and organized the literature by the access and compute each method requires and by the issues reported specifically for small models. The evidence indicates that SLMs are prone to overconfidence, that verbalized confidence is a weak signal at this scale, and that common post-training and compression steps can further degrade calibration. Our case study with three 4B-class models on medical multiple-choice questions supports a cautious but practical conclusion. Conformal prediction delivers its marginal guarantee even for a badly calibrated model, but at the cost of answering fewer than half of the questions at a 90% target, and the guarantee becomes optimistic when calibration data come from an easier subject or a different question source. Subject-conditional calibration corrects the systematic part of this problem but requires several hundred labeled examples per subject. For practitioners deploying SLMs in specialized domains, we therefore recommend calibrating on data from the deployment distribution, reporting coverage per subgroup rather than only on average, and using risk-controlling procedures when the error rate of answered questions must be bounded.

Acknowledgement

The author thanks Northeastern University for providing the GPU resources used in the experiments.

References

[1] Wang, F., Zhang, Z., Zhang, X., Wu, Z., Mo, T., Lu, Q., Wang, W., Li, R., Xu, J., Tang, X., He, Q., Ma, Y., Huang, M., & Wang, S. (2025). A comprehensive survey of small language models in the era of large language models: Techniques, enhancements, applications, collaboration with LLMs, and trustworthiness. ACM Transactions on Intelligent Systems and Technology, 16(6), 1-87. https://doi.org/10.1145/3768165

[2] Lu, Z., Li, X., Cai, D., Yi, R., Liu, F., Zhang, X., Lane, N. D., & Xu, M. (2024). Small language models: Survey, measurements, and insights. arXiv, arXiv:2409.15790. https://arxiv.org/abs/2409.15790

[3] Qwen Team. (2026). Qwen3.5: Towards native multimodal agents [Blog post]. https://qwen.ai/blog?id=qwen3.5

[4] Gemma Team. (2026). Gemma 4 technical report. arXiv, arXiv:2607.02770. https://arxiv.org/abs/2607.02770

[5] Microsoft, Abouelenin, A., Ashfaq, A., Atkinson, A., et al. (2025). Phi-4-Mini technical report: Compact yet powerful multimodal language models via mixture-of-LoRAs. arXiv, arXiv:2503.01743. https://arxiv.org/abs/2503.01743

[6] Kadavath, S., Conerly, T., Askell, A., et al. (2022). Language models (mostly) know what they know. arXiv, arXiv:2207.05221. https://arxiv.org/abs/2207.05221

[7] OpenAI, Achiam, J., Adler, S., Agarwal, S., et al. (2023). GPT-4 technical report. arXiv, arXiv:2303.08774. https://arxiv.org/abs/2303.08774

[8] Leng, J., Huang, C., Zhu, B., & Huang, J. (2025). Taming overconfidence in LLMs: Reward calibration in RLHF. In The Thirteenth International Conference on Learning Representations (ICLR 2025). https://openreview.net/forum?id=l0tg0jzsdL

[9] Zhu, C., Xu, B., Wang, Q., Zhang, Y., & Mao, Z. (2023). On the calibration of large language models and alignment. In Findings of the Association for Computational Linguistics: EMNLP 2023 (pp. 9778-9795). https://doi.org/10.18653/v1/2023.findings-emnlp.654

[10] Chen, W. (2026). ScaleCal: Scale-aware confidence calibration for small language models. Research Square preprint. https://doi.org/10.21203/rs.3.rs-10911915/v1

[11] Shen, J. (2026). Provable limits and certified deferral for verbalized uncertainty in small language models. In Multi-Disciplinary Trends in Artificial Intelligence (MIWAI 2026),

Lecture Notes in Computer Science (pp. 243-258). https://doi.org/10.1007/978-981-92-6510-7_19

[12] Shorinwa, O., Mei, Z., Lidard, J., Ren, A. Z., & Majumdar, A. (2025). A survey on uncertainty quantification of large language models: Taxonomy, open research challenges, and future directions. ACM Computing Surveys, 58(3), 1-38. https://doi.org/10.1145/3744238

[13] Liu, X., Chen, T., Da, L., Chen, C., Lin, Z., & Wei, H. (2025). Uncertainty quantification and confidence calibration in large language models: A survey. In Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining (Vol. 2, pp. 6107-6117). https://doi.org/10.1145/3711896.3736569

[14] Wen, B., Yao, J., Feng, S., Xu, C., Tsvetkov, Y., Howe, B., & Wang, L. L. (2025). Know your limits: A survey of abstention in large language models. Transactions of the Association for Computational Linguistics, 13, 529-556. https://doi.org/10.1162/tacl_a_00754

[15] Zhang, J., Cui, W., Li, Z., Huang, L., Malin, B. A., Xiong, C., & Wu, C.-S. (2026). From passive metric to active signal: The evolving role of uncertainty quantification in large language models. In Findings of the Association for Computational Linguistics: ACL 2026 (pp. 41525-41544). https://doi.org/10.18653/v1/2026.findings-acl.2064

[16] Oh, C., Park, S., Kim, T. E., Li, J., Li, W., Yeh, S., Du, S., Hassani, H., Bogdan, P., Song, D., & Li, S. (2026). Uncertainty quantification in LLM agents: Foundations, emerging challenges, and opportunities. In Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers, pp. 16219-16250). https://doi.org/10.18653/v1/2026.acl-long.738

[17] Lin, M., Lan, Q., Guan, S., Jing, W., Fan, J., Gregg, D., & Botterweck, G. (2026). Uncertainty quantification for LLM agents: A taxonomy, an evaluation protocol, and an empirical study. arXiv, arXiv:2609.07395. https://arxiv.org/abs/2609.07395

[18] Boudiaf, A., Hussain, I., & Javed, S. (2026). Uncertainty-aware decision making in multimodal large language models. arXiv, arXiv:2608.17084. https://arxiv.org/abs/2608.17084

[19] Chow, C. (1970). On optimum recognition error and reject tradeoff. IEEE Transactions on Information Theory, 16(1), 41-46. https://doi.org/10.1109/TIT.1970.1054406

[20] El-Yaniv, R., & Wiener, Y. (2010). On the foundations of noise-free selective classification. Journal of Machine Learning Research, 11(53), 1605-1641. https://jmlr.org/papers/v11/el-yaniv10a.html

[21] Geifman, Y., & El-Yaniv, R. (2017). Selective classification for deep neural networks. Advances in Neural Information Processing Systems, 30. https://papers.nips.cc/paper_files/paper/2017/hash/4a8423d5e91fda00bb7e46540e2b0cf1-Abstract.html

[22] Madras, D., Pitassi, T., & Zemel, R. (2018). Predict responsibly: Improving fairness and accuracy by learning to defer. Advances in Neural Information Processing Systems, 31. https://papers.nips.cc/paper_files/paper/2018/hash/09d37c08f7b129e96277388757530c72-Abstract.html

[23] Vovk, V., Gammerman, A., & Shafer, G. (2005). Algorithmic learning in a random world. Springer. https://doi.org/10.1007/b106715

[24] Angelopoulos, A. N., & Bates, S. (2023). Conformal prediction: A gentle introduction. Foundations and Trends in Machine Learning, 16(4), 494-591. https://doi.org/10.1561/2200000101

[25] Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., & Wasserman, L. (2018). Distribution-free predictive inference for regression. Journal of the American Statistical Association, 113(523), 1094-1111. https://doi.org/10.1080/01621459.2017.1307116

[26] Sadinle, M., Lei, J., & Wasserman, L. (2019). Least ambiguous set-valued classifiers with bounded error levels. Journal of the American Statistical Association, 114(525), 223-234. https://doi.org/10.1080/01621459.2017.1395341

[27] Campos, M. M., Farinhas, A., Zerva, C., Figueiredo, M. A. T., & Martins, A. F. T. (2024). Conformal prediction for natural language processing: A survey. Transactions of the Association for Computational Linguistics, 12, 1497-1516. https://doi.org/10.1162/tacl_a_00715

[28] Hendrycks, D., & Gimpel, K. (2017). A baseline for detecting misclassified and out-of-distribution examples in neural networks. In International Conference on Learning Representations (ICLR). https://openreview.net/forum?id=Hkg4TI9xl

[29] Jiang, Z., Araki, J., Ding, H., & Neubig, G. (2021). How can we know when language models know? On the calibration of language models for question answering. Transactions of the Association for Computational Linguistics, 9, 962-977. https://doi.org/10.1162/tacl_a_00407

[30] Wang, X., Wei, J., Schuurmans, D., Le, Q. V., Chi, E. H., Narang, S., Chowdhery, A., & Zhou, D. (2023). Self-consistency improves chain of thought reasoning in language models. In The Eleventh International Conference on Learning Representations (ICLR). https://openreview.net/forum?id=1PL1NIMMrw

[31] Kuhn, L., Gal, Y., & Farquhar, S. (2023). Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In The Eleventh International Conference on Learning Representations (ICLR). https://openreview.net/forum?id=VD-AYtP0dve

[32] Farquhar, S., Kossen, J., Kuhn, L., & Gal, Y. (2024). Detecting hallucinations in large language models using semantic entropy. Nature, 630(8017), 625-630. https://doi.org/10.1038/s41586-024-07421-0

[33] Vaithinathan Aravindan, A., & Kejriwal, M. (2026). Second guess: Detecting uncertainty through abstention and answer stability in small language models. arXiv, arXiv:2605.25394. https://arxiv.org/abs/2605.25394

[34] Lin, S., Hilton, J., & Evans, O. (2022). Teaching models to express their uncertainty in words. Transactions on Machine Learning Research. https://openreview.net/forum?id=8s8K2UZGTZ

[35] Tian, K., Mitchell, E., Zhou, A., Sharma, A., Rafailov, R., Yao, H., Finn, C., & Manning, C. D. (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 (pp. 5433-5442). https://doi.org/10.18653/v1/2023.emnlp-main.330

[36] Xiong, M., Hu, Z., Lu, X., Li, Y., Fu, J., He, J., & Hooi, B. (2024). Can LLMs express their uncertainty? An empirical evaluation of confidence elicitation in LLMs. In The Twelfth International Conference on Learning Representations (ICLR). https://openreview.net/forum?id=gjeQKFxFpZ

[37] Han, J., Buntine, W., & Shareghi, E. (2024). Towards uncertainty-aware language agent. In Findings of the Association for Computational Linguistics: ACL 2024 (pp. 6662-6685). https://doi.org/10.18653/v1/2024.findings-acl.398

[38] Azaria, A., & Mitchell, T. (2023). The internal state of an LLM knows when it's lying. In Findings of the Association for Computational Linguistics: EMNLP 2023 (pp. 967-976). https://doi.org/10.18653/v1/2023.findings-emnlp.68

[39] Kapoor, S., Gruver, N., Roberts, M., Collins, K., Pal, A., Bhatt, U., Weller, A., Dooley, S., Goldblum, M., & Wilson, A. G. (2024). Large language models must be taught to know what they don't know. Advances in Neural Information Processing Systems, 37, 85932-85972. https://papers.nips.cc/paper_files/paper/2024/hash/9c20f16b05f5e5e70fa07e2a4364b80e-Abstract-Conference.html

[40] Platt, J. C. (2000). Probabilities for SV machines. In A. J. Smola, P. Bartlett, B. Schölkopf, & D. Schuurmans (Eds.), Advances in large margin classifiers (pp. 61-74). MIT Press. https://doi.org/10.7551/mitpress/1113.003.0008

[41] Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In Proceedings of the 34th International Conference on Machine Learning (ICML), Proceedings of Machine Learning Research 70 (pp. 1321-1330). https://proceedings.mlr.press/v70/guo17a.html

[42] Zadrozny, B., & Elkan, C. (2002). Transforming classifier scores into accurate multiclass probability estimates. In Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD) (pp. 694-699). https://doi.org/10.1145/775047.775151

[43] Naeini, M. P., Cooper, G., & Hauskrecht, M. (2015). Obtaining well calibrated probabilities using Bayesian binning. Proceedings of the AAAI Conference on Artificial Intelligence, 29, 2901-2907. https://doi.org/10.1609/aaai.v29i1.9602

[44] Huang, J., Lu, P., Zeng, Q., Iwasawa, Y., Matsuo, Y., Chandar, S., Marrese-Taylor, E., & Li, I. (2026). Investigating the multilingual calibration effects of language model instruction tuning. In Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Vol. 2: Short Papers, pp. 1-59). https://doi.org/10.18653/v1/2026.eacl-short.1

[45] Proskurina, I., Brun, L., Metzler, G., & Velcin, J. (2024). When quantization affects confidence of large language models? In Findings of the Association for Computational Linguistics: NAACL 2024 (pp. 1918-1928). https://doi.org/10.18653/v1/2024.findings-naacl.124

[46] Zhang, J., Peng, X., Chen, Q., Ye, Q., Xiong, C., & Wu, C.-S. (2026). The illusion of certainty: Decoupling capability and calibration in on-policy distillation. arXiv, arXiv:2604.16830. https://arxiv.org/abs/2604.16830

[47] Meta. (2024). Llama 3.2: Revolutionizing edge AI and vision with open, customizable models [Blog post]. https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/

[48] Vovk, V., Gammerman, A., & Shafer, G. (2022). Algorithmic learning in a random world (2nd ed.). Springer. https://doi.org/10.1007/978-3-031-06649-8

[49] Romano, Y., Sesia, M., & Candès, E. J. (2020). Classification with valid and adaptive coverage. Advances in Neural Information Processing Systems, 33, 3581-3591. https://papers.nips.cc/paper_files/paper/2020/hash/244edd7e85dc81602b7615cd705545f5-Abstract.html

[50] Angelopoulos, A. N., Bates, S., Jordan, M., & Malik, J. (2021). Uncertainty sets for image classifiers using conformal prediction. In International Conference on Learning Representations (ICLR). https://openreview.net/forum?id=eNdiU_DbM9

[51] Angelopoulos, A. N., Bates, S., Fisch, A., Lei, L., & Schuster, T. (2024). Conformal risk control. In The Twelfth International Conference on Learning Representations (ICLR). https://openreview.net/forum?id=33XGfHLtZg

[52] Quach, V., Fisch, A., Schuster, T., Yala, A., Sohn, J. H., Jaakkola, T. S., & Barzilay, R. (2024). Conformal language modeling. In The Twelfth International Conference on Learning Representations (ICLR). https://openreview.net/forum?id=pzUhfQ74c5

[53] Kumar, B., Lu, C., Gupta, G., Palepu, A., Bellamy, D., Raskar, R., & Beam, A. (2023). Conformal prediction with large language models for multi-choice question answering. arXiv, arXiv:2305.18404. https://arxiv.org/abs/2305.18404

[54] Ye, F., Yang, M., Pang, J., Wang, L., Wong, D. F., Yilmaz, E., Shi, S., & Tu, Z. (2024). Benchmarking LLMs via uncertainty quantification. Advances in Neural Information Processing Systems, 37, 15356-15385. https://papers.nips.cc/paper_files/paper/2024/hash/1bdcb065d40203a00bd39831153338bb-Abstract-Datasets_and_Benchmarks_Track.html

[55] Vovk, V. (2012). Conditional validity of inductive conformal predictors. In Proceedings of the Asian Conference on Machine Learning (ACML), Proceedings of Machine Learning Research 25 (pp. 475-490). https://proceedings.mlr.press/v25/vovk12.html

[56] Tibshirani, R. J., Foygel Barber, R., Candès, E. J., & Ramdas, A. (2019). Conformal prediction under covariate shift. Advances in Neural Information Processing Systems, 32. https://papers.nips.cc/paper_files/paper/2019/hash/8fb21ee7a2207526da55a679f0332de2-Abstract.html

[57] Barber, R. F., Candès, E. J., Ramdas, A., & Tibshirani, R. J. (2023). Conformal prediction beyond exchangeability. The Annals of Statistics, 51(2), 816-845. https://doi.org/10.1214/23-AOS2276

[58] Tomani, C., Chaudhuri, K., Evtimov, I., Cremers, D., & Ibrahim, M. (2024). Uncertainty-based abstention in LLMs improves safety and reduces hallucinations. arXiv, arXiv:2404.10960. https://arxiv.org/abs/2404.10960

[59] Chen, J., Yoon, J., Ebrahimi, S., Arik, S. O., Pfister, T., & Jha, S. (2023). Adaptation with self-evaluation to improve selective prediction in LLMs. In Findings of the Association for Computational Linguistics: EMNLP 2023 (pp. 5190-5213). https://doi.org/10.18653/v1/2023.findings-emnlp.345

[60] Feng, S., Shi, W., Wang, Y., Ding, W., Balachandran, V., & Tsvetkov, Y. (2024). Don't hallucinate, abstain: Identifying LLM knowledge gaps via multi-LLM collaboration. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers, pp. 14664-14690). https://doi.org/10.18653/v1/2024.acl-long.786

[61] Ren, A. Z., Dixit, A., Bodrova, A., Singh, S., Tu, S., Brown, N., Xu, P., Takayama, L., Xia, F., Varley, J., Xu, Z., Sadigh, D., Zeng, A., & Majumdar, A. (2023). Robots that ask for help: Uncertainty alignment for large language model planners. In Proceedings of the 7th Conference on Robot Learning, Proceedings of Machine Learning Research 229 (pp. 661-682). https://arxiv.org/abs/2307.01928

[62] Chen, L., Zaharia, M., & Zou, J. (2024). FrugalGPT: How to use large language models while reducing cost and improving performance. Transactions on Machine Learning Research. https://openreview.net/forum?id=cSimKw5p6R

[63] Ding, D., Mallick, A., Wang, C., Sim, R., Mukherjee, S., Rühle, V., Lakshmanan, L. V. S., & Awadallah, A. H. (2024). Hybrid LLM: Cost-efficient and quality-aware query routing. In The Twelfth International Conference on Learning Representations (ICLR 2024). https://openreview.net/forum?id=02f3mUtqnM

[64] Ong, I., Almahairi, A., Wu, V., Chiang, W.-L., Wu, T., Gonzalez, J. E., Kadous, M. W., & Stoica, I. (2025). RouteLLM: Learning to route LLMs from preference data. In The Thirteenth International Conference on Learning Representations (ICLR 2025). https://openreview.net/forum?id=8sSqNntaMr

[65] Gupta, N., Narasimhan, H., Jitkrittum, W., Rawat, A. S., Menon, A. K., & Kumar, S. (2024). Language model cascades:

Token-level uncertainty and beyond. In The Twelfth International Conference on Learning Representations (ICLR 2024). https://openreview.net/forum?id=KgaBScZ4VI

[66] Chuang, Y.-N., Yu, L., Wang, G., Zhang, L., Liu, Z., Cai, X., Sui, Y., Braverman, V., & Hu, X. (2025). Confident or seek stronger: Exploring uncertainty-based on-device LLM routing from benchmarking to generalization. arXiv, arXiv:2502.04428. https://arxiv.org/abs/2502.04428

[67] Davidov, H., Cohen, N., Kalinsky, O., Fairstein, Y., Kushilevitz, G., Yazdi, R., & Rebeschini, P. (2026). Knowing when to quit: A principled framework for dynamic abstention in LLM reasoning. In Proceedings of the 43rd International Conference on Machine Learning, Proceedings of Machine Learning Research 306 (pp. 23121-23158). https://proceedings.mlr.press/v306/davidov26a.html

[68] Luo, H., Wen, B., & Wang, L. L. (2026). Agentic abstention: Do agents know when to stop instead of act? arXiv, arXiv:2606.28733. https://arxiv.org/abs/2606.28733

[69] Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. Monthly Weather Review, 78(1), 1-3. https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2

[70] Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2021). Measuring massive multitask language understanding. In International Conference on Learning Representations (ICLR). https://openreview.net/forum?id=d7KBjmI3GmQ

[71] Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C., & Tafjord, O. (2018). Think you have solved question answering? Try ARC, the AI2 reasoning challenge. arXiv, arXiv:1803.05457. https://arxiv.org/abs/1803.05457

[72] Jin, D., Pan, E., Oufattole, N., Weng, W.-H., Fang, H., & Szolovits, P. (2021). What disease does this patient have? A large-scale open domain question answering dataset from medical exams. Applied Sciences, 11(14), 6421. https://doi.org/10.3390/app11146421

[73] Pal, A., Umapathi, L. K., & Sankarasubbu, M. (2022). MedMCQA: A large-scale multi-subject multi-choice dataset for medical domain question answering. In Proceedings of the Conference on Health, Inference, and Learning (CHIL), Proceedings of Machine Learning Research 174 (pp. 248-260). https://proceedings.mlr.press/v174/pal22a.html

[74] Lin, S., Hilton, J., & Evans, O. (2022). TruthfulQA: Measuring how models mimic human falsehoods. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers, pp. 3214-3252). https://doi.org/10.18653/v1/2022.acl-long.229

[75] Fadeeva, E., Vashurin, R., Tsvigun, A., Vazhentsev, A., Petrakov, S., Fedyanin, K., Vasilev, D., Goncharova, E., Panchenko, A., Panov, M., Baldwin, T., & Shelmanov, A. (2023). LM-Polygraph: Uncertainty estimation for language models. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (pp. 446-461). https://doi.org/10.18653/v1/2023.emnlp-demo.41

[76] Vashurin, R., Fadeeva, E., Vazhentsev, A., Rvanova, L., Vasilev, D., Tsvigun, A., Petrakov, S., Xing, R., Sadallah, A., Grishchenkov, K., Panchenko, A., Baldwin, T., Nakov, P., Panov, M., & Shelmanov, A. (2025). Benchmarking uncertainty quantification methods for large language models with LM-Polygraph. Transactions of the Association for Computational Linguistics, 13, 220-248. https://doi.org/10.1162/tacl_a_00737