PROCEEDINGS A

royalsocietypublishing.org/journal/rspa

Research

Check for updates

Article submitted to journal

Subject Areas:
machine learning, natural language processing, large language models, applied statistics

Keywords:
large language models, uncertainty quantification, conformal prediction

Author for correspondence:
Alice Evelyn Ashby
e-mail: Alice.Ashby.2025@live.rhul.ac.uk

Uncertainty-Aware Large Language Models: A Scoping Review of Conformal Prediction Methods

Alice Evelyn Ashby[1,2], Khuong An Nguyen[1] and Zhiyuan Luo[1]

[1]Centre for Reliable Machine Learning, Department of Computer Science, Royal Holloway University of London, UK

[2]Surrey Institute for People-Centred AI, University of Surrey, UK

In recent years, there has been growing interest in applying Conformal Prediction to large language models (LLMs) across diverse domains to enhance the trustworthiness of their predictions. However, the literature still lacks a comprehensive survey of this rapidly emerging area. Thus, to fill this gap, this article presents a comprehensive review and analysis of conformal prediction for LLMs. We review over 106 studies, and propose a novel taxonomy that categorises existing methods into six groups. In addition, we observe trends in the literature with regard to LLM, dataset, and task selection, and make recommendations for researchers to improve task diversity and address gaps in black-box LLM uncertainty estimation.

Interestingly, we find that logit-free methods tend to outperform logit-based methods, with a lower mean absolute coverage error and prediction set size. This suggests that logit-free methods, which rely on uncertainty signals based on self-consistency sampling, semantic diversity, and other methods, may sidestep known issues with mis-calibrated token probabilities and may have advantageous robustness to tokenisation and decoding idiosyncrasies, particularly for open-ended generation where the performance gap is more pronounced. This indicates that black-box uncertainty signals may more directly capture semantic correctness or answer stability.

We also find that on average, conformal methods for Large Vision Language Models (LVLMs) have higher over-coverage error than LLMs, and almost non-existent under-coverage error, suggesting that methods for LVLMs may be more conservative. While we do not claim a definitive causal explanation, empirical evidence suggests that conformal methods for LVLMs exhibit a stronger coverage-informativeness trade-off than those for LLMs.

THE ROYAL SOCIETY PUBLISHING

1. Introduction

In recent years, there has been a surge in large language model (LLM) deployments across a broad range of natural language processing (NLP) tasks in various critical domains, including medicine [1], robotics [2], and law [3], where their scale and pre-training on massive corpora give them strong generative and reasoning capabilities. However, LLMs are not infallible; they pose serious reliability concerns in these high-stakes domains, due to their propensity to produce overconfident, hallucinated, biased, or otherwise unsafe outputs [4–8]. Conformal prediction is an uncertainty quantification paradigm that provides a statistical layer for addressing these concerns; as a model-agnostic, distribution-free calibration framework with finite-sample guarantees [9–12], it can wrap black-box predictors to estimate uncertainty without training and thereby make downstream decisions more trustworthy. It has been successfully deployed in several high-stakes settings [13–19], and in this paper, we demonstrate the success of conformal prediction in the LLM setting. Conformal prediction improves the reliability and safety of LLMs by controlling risks such as hallucination and toxicity, while also supporting more efficient deployment through techniques such as routing and token-level calibration, which reduce inference cost and improve token efficiency. Through calibrating prediction sets, selective prediction with abstention and deferral, and reliability guarantees for diverse outputs and information-retrieval mechanisms, conformal prediction can improve the practical accuracy of LLMs; not by modifying the base models themselves, but by ensuring that accepted outputs are more likely to be correct, factual, and aligned with human values and end user requirements.

(a) Review Objectives

We intend to provide a detailed over-arching overview of the field of conformal prediction LLM uncertainty quantification. All conformal methods we identify and that meet our inclusion criteria will be categorised in a hierarchical taxonomy for comprehensive stratification and subsequently, ease-of-navigation for academic scholars and industry practitioners interested in this research area. We conduct an in-depth and extensive analysis of the performance of the categorised conformal methods, as well as report on trends we observe in the literature.

- **Hypothesis A.** An LLM's pre-softmax outputs are called logits (see Section 3 for more details). In general, conformal methods that exploit logits, or quantities derived from them such as log-probabilities, are expected to produce stronger uncertainty estimates, since logits provide a richer and more direct view of a model's confidence than the generated output alone. In practice, however, such information is often unavailable for proprietary closed-source, API-based LLMs, such as OpenAI's ChatGPT. This has motivated the development of logit-free conformal methods that rely on external signals, such as reprompting, self-consistency sampling, or entailment scores. Whilst these approaches are attractive because they apply to black-box models, **we hypothesise that the resulting uncertainty signal will be noisier than those obtained from logits. In conformal terms, we expect to see a higher coverage error and larger and less informative prediction sets when using logit-free methods.**
- **Hypothesis B.** Large vision language models (LVLMs) extend the generation and reasoning capabilities of LLMs to multimodal inputs by combining a visual encoder with a language model backbone. LVLMs support multimodal tasks such as scene and document understanding or visual question-answering, but these settings often introduce additional sources of uncertainty arising from visual perception, cross-modal fusion, and ambiguity in grounding text to image regions. As a result, **we hypothesise that conformal methods for LVLMs may face a more challenging calibration problem than the text-only setting of LLMs due to these additional sources of uncertainty. In conformal terms, we expect to see higher coverage error, particularly overcoverage error, and larger and less informative prediction sets in vision-language tasks**.

(b) Our Contributions

Since its inception mid-2023, the field has seen a rise in publications, as illustrated in Figure 1. This appears to follow a linear trend, projected to continue based on the current number of papers published in 2026. However, there is also the possibility of exponential growth, as the number of publications in 2024 doubled in 2025. Interestingly, the majority of publications are freely available, signifying a systemic shift with preprints identified as a key contributor.

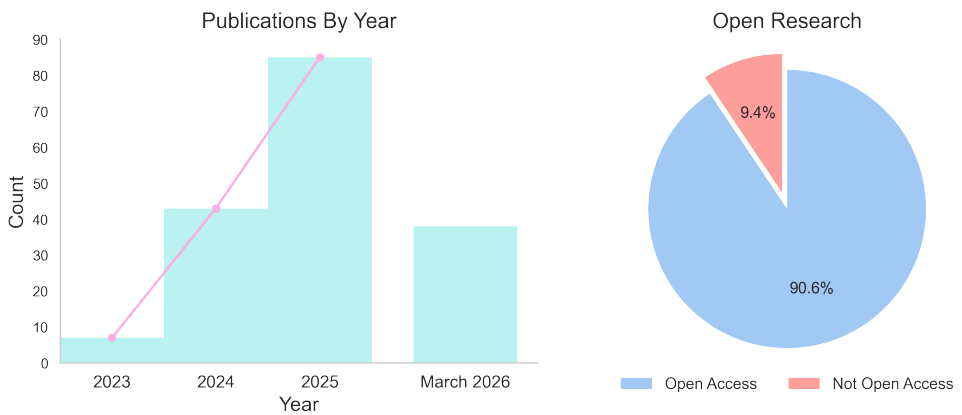

Figure 1: The number of papers published by year and their open access status, according to OpenAlex data. We used the search terms "`conformal prediction large langu--age models`" with date range from 2023 to present (March 2026), searching on title and abstract with stemming enabled, excluding any non-article, non-preprint, non-dissertation.

However, **despite the evident growth of conformal prediction in the natural language processing field, a dedicated scoping review has yet to be conducted.** To address this urgent need, this article presents the first comprehensive review and empirical analysis of conformal prediction for uncertainty-aware LLMs. We make the following contributions:

- We conduct the **first in-depth and extensive analysis of 106 papers on conformal prediction applied to LLMs**, covering the full state-of-the-art from the inception of this field to the early months of 2026. We also provide a GitHub repository [20] as a structured reference for ongoing conformal prediction research for LLMs.
- We present a **novel taxonomy of conformal methods for LLM uncertainty quantification** to represent the current landscape of research and practise. Specifically, we categorise all existing methods into 6 groups: open-ended conformal calibration, close-ended conformal calibration, conformal selective prediction, conformal retrieval-augmented generation, conformal factuality, and conformal abstention.
- We conduct a **comprehensive controlled performance analysis of 30 methods across our 6 taxonomy categories**, on a diverse selection of tasks, datasets, user-specified error rates, and settings. We then identify the most effective methods across these heterogeneous settings, and determine where gaps in the research landscape yet to be addressed.
- We **systematically synthesise key characteristics and quantify the performances** of conformal methods designed for or applied to LLMs. Specifically, we examine: (1) the language models utilised, (2) the tasks or applications explored, (3) the frameworks the proposed methods are based on if applicable, (4) the datasets the proposed method is evaluated on, (5) the evaluation metrics utilised.

The remainder of this paper is structured as follows. Section 2 introduces the review focus and research methodology adopted by this paper, as well as a comparison to previous surveys. Section 3 provides a preliminary on large language models for readers not familiar with the field, Section 4 introduces conformal prediction for readers new to the framework. Section 5 outlines our proposed taxonomy of conformal methods for LLMs and details our performance analyses. Section 6 examines the trends in the literature. Sections 7, 8, 9, 10, 11, and 12 survey the literature, and finally, Section 13 concludes the paper.

2. Review Scope and Methodology

To ensure the reproducibility of the literature selection process, this review adopts a systematic methodology inspired by the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [21]. The PRISMA framework follows a structured four-phase process: **(1) Identification** (searching databases with predefined keywords), **(2) Screening** (filtering records based on titles/abstracts), **(3) Eligibility** (assessing fulltext articles against inclusion/exclusion criteria), and **(4) Inclusion** (finalising selected studies). This section outlines the research scope in detail, provides the databases, keywords, and inclusion/exclusion criteria utilised, and describes the systematic approach used to search for, select, screen, include, exclude, and analyse existing conformal prediction methods for LLM and LVLM applications, thereby ensuring a rigorous and comprehensive examination on the topic.

(a) Review Focus and Selection Criteria

This paper is dedicated to providing a comprehensive review and in-depth analysis of current conformal prediction methods applied to LLMs, with the goal of offering valuable and unique insights into the trends and challenges in this field. We performed an extensive search using the keyword pairs from group A combined with keyword pairs from group B, where $A =$ {"`conformal prediction`", "`conformal inference`", "`conformal risk control`", "`conformalised selection`"} and $B =$ {"`large language models`", "`small language models`", "`multimodal large language models`", "`large vision language models`"}. If we choose $a_i \in A$ and $b_j \in B$, and define $s_{ij} = a_i + b_j$. There are 16 possible different s_{ij}. Therefore, $S = \{s_{ij} : 1 \leq i \leq 4, 1 \leq j \leq 4\}$.

Each string $s \in S$ is used as input to the OpenAlex structured knowledge graph, using boolean search in the title and abstracts only, with stemming enabled. We found that searching the title, abstract, and fulltext identified too many out-of-scope papers. We searched specifically for research articles, dissertations, and preprints; the inclusion of preprints was decided due to them being a popular medium for the rapid dissemination of foundational research, especially in fast growing research landscapes like that of LLMs. We included dissertations as graduate students may have contributed to the topic. Initially, we searched a date range from 2020 to the present, but we found no relevant papers pre-2023.

Using Google Scholar, we ran independent searches to isolate preprints and dissertations. For preprints, we added logical operators to our previous notation, such that a search string is created like so, ["`conformal prediction`" `AND` "`large language models`"] `site:arxiv.org`. This returned only preprints from the arXiv server, which we consider the most reliable. For dissertations, we use this format but target ProQuest in place of arXiv. ProQuest is a premier and reputable database of graduate-level research. The search string looked like so, ["`conformal prediction`" `AND` "`large language models`"] `site:proquest.com`. For both searches, we search with a date range of 2023 to present.

The inclusion criteria for studies in this review were as follows:

- Peer-reviewed articles, arXiv preprints, or dissertations published in English.
- Studies must adapt an existing conformal framework or propose a novel conformal prediction (CP) method. Acceptable frameworks may include split or full CP, conformal

risk control, conformal selection, adaptive conformal inference, conformalised quantile regression, conformal Venn-ABERS, or learn-then-test. Studies that provide PAC guarantees may be included if the guarantee is provided with learn-then-test, or a CP variant such as PAC conformal prediction. Benchmarking studies that perform large-scale analysis, for example evaluating several LLMs, tasks, and nonconformity scoring functions are acceptable.

- Studies must include at least one application that utilises either a decoder-only architecture, such as generative pre-trained transformer (GPT), or encoder-decoder architecture, such as text-to-text transfer transformer (T5). LLMs with parameter counts in the range of 60 million (M) to 70 billion (B), or even 1 trillion (T) for Mixture of Experts (MoE) models are acceptable; a full scaling classification is provided in Section 5. LLMs with a pre-trained vision encoder, termed large vision language models (LVLMs), are acceptable provided that they have an autoregressive text decoder as with GPT and T5.

Exclusion criteria encompassed literature in any other language and studies that did not exceed an internal quality threshold we set for screening. Furthermore, in accordance with the criteria and our research scope, studies that only utilise encoder-only architectures such as bidirectional encoder representations from transformers (BERT) or encoder-only foundation models such as contrastive language-image pre-training (CLIP) or grounding DINO are excluded.

Following this methodology, a comprehensive literature search was conducted across OpenAlex and Google Scholar using the specified keywords. At this initial stage, **we had identified 488 papers, which were reduced to 165 papers** after the removal of duplicates and the elimination of less relevant studies through title and abstract vetting. We then assessed eligibility with a meticulous manual review and critical analysis of each paper's technical content, methodological approach, and experimental validation, and in particular its applicability to LLMs as per our inclusion criteria. The selection was thus refined to **a final set of 106 publications**.

(b) Comparison to Existing Surveys

Recent surveys have considered CP in natural language processing (NLP), CP across broader ml settings, or uncertainty quantification for LLMs. However, our review stands out as the **first to provide a dedicated synthesis of conformal prediction for LLMs as a distinct research area, with over 106 studies**. None of the existing reviews matches ours in terms of specificity, coverage, and synthesis depth. We explore the differences between our review and these existing reviews.

- *Conformal Prediction for Natural Language Processing: A Survey* [22]
  A broad survey of CP in NLP. It includes some studies of CP for LLMs, but this is only a small selection in the wider NLP landscape. The review only covers works from before May 2024, and so many recent studies are excluded (see Figure 1).
- *Conformal Prediction: A Data Perspective* [23]
  A general CP survey organised by datatype and proposes a data-centric taxonomy. Some LLM-adjacent CP work is selectively covered, but this is very limited.
- *A Survey on Uncertainty Quantification of Large Language Models* [24]
  A broad general survey on uncertainty quantification for LLMs, with a taxonomy organised around token-level, self-verbalised, semantic-similarity, and mechanistic-interpretability. CP is only briefly discussed in the calibration section.

3. Preliminary: Large Language Models

LLMs are neural sequence models trained to assign probabilities to text, typically by predicting the next token given a preceding context. Formally, if $x_{1:N}$ denotes a tokenised input sequence, an LLM models the conditional distribution $\mathbb{P}_\theta(x_n|x_{<n})$ which is a learned parameterised mapping from the context $x_{<n}$ to the next token x_n. Text generation then proceeds autoregressively by

repeatedly sampling or selecting from this distribution [25]. The LLM outputs a vector of logits $z_v(x) \in \mathbb{R}^{|\mathcal{V}|}$ for the context x over the vocabulary $\mathcal{V}$, for each $v \in \mathcal{V}$, which is converted into probabilities by the softmax map,

$$\mathbb{P}_\theta(v \mid x) = \frac{\exp(z_v(x))}{\sum_{u \in \mathcal{V}} \exp(z_u(x))}, \tag{3.1}$$

and then selects or samples one token from these probabilities. The logarithms of these probabilities are the log-probabilities (logprobs), which are often exposed by model APIs and have a central role in uncertainty quantification, decoding, and downstream calibration. A high-level view, therefore, is that LLMs are probabilistic predictors over token sequences, even when deployed as conversational (where the model is often suffixed with `Chat`) or instruction-tuned (where the model is often suffixed with `Instruct`).

Most contemporary LLMs are based on the decoder-only transformer architecture, which combines token embeddings, positional encodings, masked self-attention, and feed-forward blocks to produce contextualised representations of the input prefix [25]. Self-attention allows each token to interact with all earlier tokens, thereby capturing long-range dependencies more effectively than earlier recurrent architectures. This architecture underlies the widespread pre-training paradigm; large-scale self-supervised training on massive corpora using the autoregressive language-modelling objective, $\mathcal{L}_{\mathrm{LM}}(\theta) = -\sum_{n=1}^{N} \log \mathbb{P}_\theta(x_n \mid x_{<n})$, followed by adaptation through prompting, supervised fine-tuning, or preference optimisation [25]. In practice, the raw probabilistic outputs of an LLM are often miscalibrated (see Section ii for more information), particularly after reinforcement learning with human feedback (RLHF), instruction-finetuning, or alignment. A standard post-hoc correction is temperature scaling (TS), which uses a scalar $T > 0$ and applies softmax to the rescaled logit vector $z_v(x)/T$,

$$\mathbb{P}_T(v \mid x) = \mathrm{softmax}(z_v(x)/T) = \frac{\exp(z_v(x)/T)}{\sum_{u \in \mathcal{V}} \exp(z_u(x)/T)}, \tag{3.2}$$

thereby sharpening or flattening the predictive distribution without changing the ranking of candidate tokens [26]. Prompting is the principal interface through which pre-trained LLMs are adapted at inference time. Rather than updating parameters, one specifies a textual template that conditions the model toward a desired task, style, or reasoning pattern. This includes zero-shot prompting, few-shot in-context learning, and chain-of-thought prompting, all of which exploit an LLM's sensitivity to context while leaving the model weights fixed [25].

Below are some additional points which readers may find relevant to the rest of our paper.

- **Encoder-Decoder Transformers.** In this architecture, the encoder maps the input sequence to contextual representations, and the decoder generates the output sequence conditioned on those representations. They are most often used in sequence-to-sequence tasks. An example is the text-to-text transfer transformer (T5), which casts a wide range of tasks to this format, i.e. translation, summarisation, structured text transformation.
- **Mixture of Experts.** Architecturally, LLMs may be either dense or mixture-of-experts (MoE) models. In dense models, all parameters are activated for every token. In contrast, MoE architectures contain multiple expert subnetworks together with a routing mechanism that activates only a subset of experts per token, thereby increasing total parameter count without proportionally increasing inference cost.
- **Large Vision Language Models.** Or LVLMs, extend the generation and reasoning capabilities of LLMs to multimodal inputs by combining a visual encoder with a language model backbone. For example, LLaVA-1.5 couples a contrastive language-image pre-training (CLIP) visual encoder with a Vicuna LLM backbone through a two-layer MLP-based vision-language connector [27]. In systems such as LLaVA and others, images are converted into embeddings and fused with textual tokens such that the model can condition on both modalities when generating a response. That means conceptually,

LVLMs inherit the probabilistic next-token prediction of LLMs, but their uncertainty now jointly depends on linguistic and visual data.

- **Small Language Models.** Compact variants of language models designed to operate under tighter computational and memory constraints than frontier-scale LLMs. Although they typically underperform the largest models in absolute capability, SLMs are cheaper to train and deploy, especially for latency-sensitive or resource-constrained settings.
- **Open vs. Closed-source Models.** Open-source models provide public model weights and local control over inference, enabling direct access to logits, hidden states, and other internal components. Popular open-source model families include LLaMA, Qwen, and DeepSeek. Closed-source models are generally accessed via an API and may expose only partial information, such as sampled outputs, token logprobs, or confidence surrogates, though not always. Closed-source models include GPT-5, Claude Opus, and Gemini.

Much remains to be understood about LLMs; interested readers are referred to [25] for a comprehensive overview.

4. Preliminary: Conformal Prediction

CP is a model-agnostic framework for uncertainty quantification that converts a pre-trained predictor into set-valued predictions with finite-sample, distribution-free validity under exchangeability [9–12,28]. Its central objective is to produce prediction sets that are informative while guaranteeing that the true outcome is included with user-specified coverage probability $1-\alpha$, regardless of model correctness or distributional form. As such, CP can be considered a lightweight post-hoc calibration layer for black-box predictors, and is particularly valuable in high-stakes settings that appreciate its explicit and non-asymptotic guarantees, which do not require retraining the underlying model. In this preliminary, we focus on **split conformal prediction (SCP)** in the classification setting, providing a concise overview of the framework.

Let $\hat{f}(x) \in [0,1]^K$ denote a trained classifier returning class scores or probabilities for K possible labels, and let $\{(x_i, y_i)\}_{i=1}^n$ be an i.i.d.[1] calibration set of n examples where $x_i \in \mathcal{X}$ and $y_i \in \mathcal{Y}$, independent of model training. Given a nonconformity score $s(x,y)$ and a miscoverage rate α, one computes calibration scores $s_i = s(x_i, y_i)$ and their adjusted empirical quantile $\hat{q} = \text{Quantile}\left(\{s_i\}_{i=1}^n ; \frac{\lceil (n+1)(1-\alpha) \rceil}{n}\right)$. The conformal prediction set for a new input x_{n+1} is then $C(x_{n+1}) = \{\, y : s(x_{n+1}, y) \le \hat{q} \}$. Under exchangeability, this construction satisfies the marginal coverage guarantee $\mathbb{P}(y_{n+1} \in C(x_{n+1})) \ge 1-\alpha$. While natural language is inherently sequential and non-exchangeable, the exchangeability assumption is a mathematical approximation that unlocks powerful uncertainty quantification and statistical validity for LLM pipelines.

A nonconformity score is a function that measures how unusual an example is relative to a trained model and previously seen data. It is the core ingredient in conformal prediction. A larger score indicates that the model is less confident or that the data point appears more anomalous. The choice of nonconformity score function determines the efficiency and adaptivity of the resulting prediction sets, while the validity guarantee is inherited from the conformal calibration step [10].

For classification, there are several nonconformity score functions that have now become standard. The least ambiguous set-valued classifier (LAC), for example, uses the score, $s_{\text{LAC}}(x,y) = 1 - \hat{\pi}_y(x)$, where $\hat{\pi}_y(x)$ is the model's estimated probability of class y for x. Thus, labels are included according to their individual predicted probabilities alone [29]. LAC is a simple score that is easy to interpret, but it lacks adaptivity as it does not account for the overall shape of the predictive distribution, making it less efficient in difficult multiclass settings. Adaptive prediction sets (APS), however, define the nonconformity score of a candidate label y by the cumulative predicted probability mass required to reach that label in the model's sorted output distribution, $s_{\text{APS}}(x,y) = \rho_x(y) + \hat{\pi}_y(x)\, u$, where $\rho_x(y)$ is the total probability mass of labels ranked above y, and $u \sim \text{Unif}(0,1)$ is a randomisation term [30]. Intuitively, APS

[1] Independent and identically distributed.

includes labels in decreasing order of predicted probability until sufficient cumulative mass has been accumulated, thereby yielding prediction sets that adapt to the shape of the predictive distribution. Regularised adaptive prediction sets (RAPS) augment APS with a rank-based regularisation term, $s_{\mathrm{RAPS}}(x,y) = \rho_x(y) + \hat{\pi}_y(x)\,u + \lambda\big(\sigma_x(y) - k_{\mathrm{reg}}\big)_+$, where $(z)_+$ represents the positive part of z, $\sigma_x(y)$ denotes the rank of label y, k_{reg} is an unpenalised rank threshold, and $\lambda \geq 0$ controls the strength of the penalty [31]. Therefore, $\lambda = 0$ recovers APS. This additional term discourages the inclusion of labels deep in the ranked list, whose small tail probabilities are often unstable, and therefore tends to produce smaller and more stable prediction sets.

Beyond miscoverage control, **conformal risk control (CRC)** generalises CP to arbitrary monotone losses [32]. Instead of controlling the event $y \notin C(x)$, CRC considers a family of set-valued predictors C_λ indexed by a conservativeness parameter λ, together with a bounded loss $\mathcal{L}(C_\lambda(x), y)$ that is non-increasing in λ. The target becomes $\mathbb{E}\big[\mathcal{L}\big(C_{\hat{\lambda}}(x_{n+1}), y_{n+1}\big)\big] \leq \alpha$. Writing $\mathcal{L}_i(\lambda) = \mathcal{L}(C_\lambda(x_i), y_i)$ and $\widehat{R}_n(\lambda) = n^{-1}\sum_{i=1}^{n}\mathcal{L}_i(\lambda)$, $\hat{\lambda} = \inf\left\{\lambda : \frac{n}{n+1}\widehat{R}_n(\lambda) + \frac{B}{n+1} \leq \alpha\right\}$ is selected by CRC, where B is an upper bound on the loss. This subsumes CP when $\mathcal{L}$ is the miscoverage indicator, but also accommodates more task-relevant notions of error through structured risk, such as token-level F-measure loss, or false negative rate (FNR).

Quantile risk control (QRC) extends the focus from expected loss to quantiles and more general quantile-based risk measures of the loss distribution [33]. Rather than bounding only the mean loss, it constructs confidence bounds on the loss cumulative distribution function or equivalently on the quantile function $F^{-1}(p) = \inf\{x : F(x) \geq p\}$. These bounds can then be integrated against a weighting function $\psi(p)$ to control any quantile-based risk measure, $R_\psi(F) = \int_0^1 \psi(p)\,F^{-1}(p)\,dp$, where $\psi(p) \geq 0$ and $\int_0^1 \psi(p)\,dp = 1$. The key idea is to move beyond standard accuracy-style guarantees and instead directly control the tail behaviour of the loss distribution. This class includes the value-at-risk, conditional value-at-risk, and related tail risks.

Several important generalisations, variants, and adjacent frameworks to conformal prediction, which are relevant to the contents of this review, are summarised below.

Conformal Selection (CS). A framework for screening problems in which one wishes to select test points whose unobserved outcomes exceed user-specified thresholds while controlling false discoveries [34,35]. The method constructs conformal p-values that quantify evidence that a test outcome is 'large', and then applies a multi-step testing rule such as Benjamini-Hochberg, to obtain a selection set with finite-sample false discovery rate control.

Learn-then-Test (LTT). A framework that reframes risk control as a multiple hypothesis testing problem over candidate parameter values, and then uses family-wise error rate correction to identify settings whose risk can be certified below the target level [36]. LTT is useful when the risk is non-monotone or when multiple interacting risks must be controlled.

Risk-controlling Prediction Sets (RCPS). A framework that provides a high-probability alternative to CRC when one wishes to control a general monotone loss through set-valued predictions with probability at least $1 - \delta$ [37], where δ is the user-specified error level. The framework begins with a nested family of set predictors, a loss function on the prediction set that decreases as the prediction set grows, and a risk of a set-valued predictor. RCPS calibrates the parameter λ by constructing an upper confidence bound on the risk tolerance level α and selecting the smallest $\hat{\lambda}$ such that all larger λ values have upper confidence bounds below the target level α. RCPS is typically more conservative and depends on both the risk tolerance level α and the error level δ.

Conditional Conformal Prediction. The quasi-conditional conformal prediction (QCCP) framework by Gibbs et al. aims to interpolate between marginal and fully conditional validity by requiring coverage over a user-specified class of covariate shifts or weighting functions [38]. Exact finite-sample conditional coverage is impossible [39], but this framework provides principled finite-sample guarantees over structured families of shifts, thereby sharpening marginal coverage protection. By contrast, Vovk's conditional SCP framework [40] formulates conditional validity through conditioning on different elements of the prediction problem, such as the training set,

the test object, the label, or combinations thereof, and shows that SCP can be modified to achieve finite-sample guarantees for certain group-based notions of conditional validity. The Mondrian class-conditional framework calibrates separately on classes, yielding guarantees in the form, $\mathbb{P}(y_{n+1} \in C(x_{n+1}) \mid y_{n+1} = y) \geq 1 - \alpha$, for all $y \in \mathcal{Y}$, at the cost of a reduced effective calibration sample size within each class [41].

Weighted Conformal Prediction. Tibshirani et al. extend validity beyond exchangeability to covariate shift by replacing the empirical quantile of calibration scores with a weighted quantile based on estimated density ratios between test and calibration covariate distributions [42], whereas Barber et al. address more general non-exchangeable settings through fixed weighted conformal procedures and coverage-gap bounds that also permit nonsymmetric algorithms [43].

Conformal Quantile Regression (CQR). A framework for conformal regression, when the goal is to obtain prediction intervals that are both distribution-free and adaptive to heteroscedasticity [44]. The usual conformal residual correction around a conditional mean estimate is replaced with a plug-in interval formed from estimated lower and upper conditional quantiles, and then this interval is conformalised on a held-out calibration set. The finite-sample marginal coverage guarantee of SCP is inherited, now with intervals for local uncertainty.

Other generalisations and variants. These include CP with e-values in [45], CP for PAC learning [46] in [47], adaptive conformal inference (ACI) for distribution shift in [48], and statistical dispersion control for distribution-free population-level risk control [49].

That concludes our preliminary on conformal prediction. Interested readers are encouraged to explore these topics in more depth in [12], or explore the abundance of resources found online[2].

5. Taxonomy of Conformal Methods for Large Language Models

We propose a hierarchical tree-structured taxonomy of conformal methods applied to the field of LLMs. In this taxonomy, the foundation is the root nodes, which correspond to frameworks in the conformal prediction family outlined in Section 4.

Figure 2 shows that the primary root node is the SCP framework, which extends to the conformal selection framework and generalises to the CRC framework, which is adjacent/closely related to the LTT framework. Together, these root nodes feed into several branch nodes which apply a conformal framework to a specific task or application in the LLM space.

Below we define the branch nodes:

- **Open-ended Conformal Calibration** Conformal methods based on either SCP, CRC, or LTT frameworks that operate in the unbounded output space of LLMs, such as text generation tasks like open question-answering and summarisation.
- **Close-ended Conformal Calibration** Conformal methods based on either SCP or CRC frameworks that operate in the constrained output space of LLMs, such as text classification tasks and multiple-choice question-answering.
- **Conformal Selective Prediction** Conformal methods based on either SCP, CS, CRC, or LTT frameworks that focus on providing admissible candidate LLM responses to user prompts, abstaining or deferring to another LLM if unsure.
- **Conformal Retrieval-augmented Generation** Conformal methods based on either SCP, CRC, or LTT frameworks that provide reliability guarantees to RAG systems, which reduce hallucinations by grounding LLM responses with external knowledge base data.
- **Conformal Factuality** Conformal methods based on either SCP or LTT frameworks that emphasise the correctness of LLM generations, by ensuring each sub-claim of a long-form response is factually accurate, thereby reducing hallucinations.

[2] Available at: https://github.com/valeman/awesome-conformal-prediction.

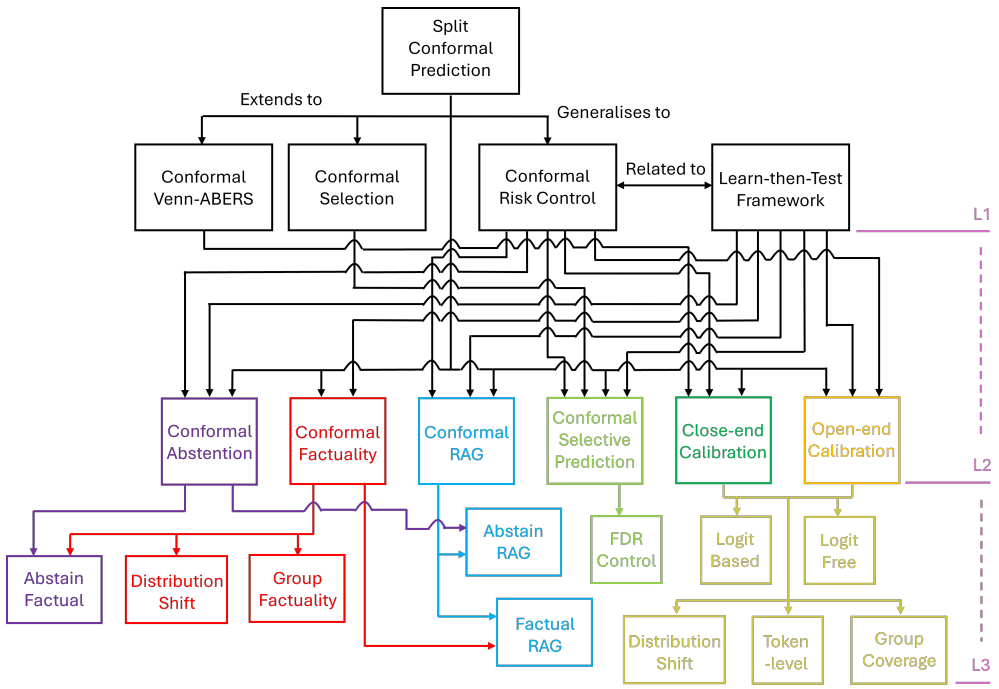

Figure 2: A hierarchical tree-structured taxonomy of conformal methods applied to LLMs. The tree has three distinct levels: L1—the root nodes, which are generalisations or extensions of the split conformal prediction framework that serve as a foundation for the following; L2—the branch nodes, which apply these conformal frameworks to various specific LLM tasks or applications; and L3—the leaf nodes, which further decompose these categories into specific related subcategories. The branch and leaf nodes are colour-coded accordingly in Section 5.

- **Conformal Abstention** Conformal methods based on either SCP, CRC, or LTT frameworks that will abstain from answering queries that they do not know the answer to, thereby reducing hallucinations.

Branch nodes can have associated leaf nodes which further decompose these categories into related subcategories; the open-ended and close-ended conformal calibration nodes have shared leaf nodes, and other branch nodes can have leaf nodes with functional overlap.

Below we define the leaf nodes:

- **Logit-based Conformal Calibration** Conformal methods that rely on logits/logprobs to calibrate LLMs. If conformal methods also require internal embeddings or other representations such as hidden states, attention features, but not necessarily logits or equivalent, we still place them in this leaf node. It can therefore be alternatively considered exclusively white-box LLM calibration.
- **Logit-free Conformal Calibration** Conformal methods that do not rely on logits/logprobs or internal embeddings and equivalent representations to calibrate LLMs. It can therefore be alternatively considered exclusively true black-box LLM calibration.
- **Token-level Conformal Calibration** Conformal methods that calibrate the next-token distribution at each autoregressive step in LLM generation, with prediction sets over tokens rather than full sequences, unlike sequence-level conformal methods.
- **Group-conditional Conformal Calibration** Conformal methods that provide group-conditional coverage guarantees rather than simple marginal coverage guarantees.

- **Distribution-shift-aware Conformal Calibration** Conformal methods that provide coverage guarantees robust to different types of distribution shift, such as covariate shift.
- **Conformal Selective Prediction with FDR Control** Conformal methods that select a subset of instances for which the FDR among selections is rigorously controlled.
- **Conformal Factuality with Retrieval-augmented Generation** Conformal methods for calibrating LLM retrieval mechanisms, where guarantees on the factuality of retrieved data from the external knowledge base are provided.
- **Abstention-aware Retrieval-augmented Generation** Conformal methods for calibrating LLM retrieval mechanisms with the ability to abstain from answering user queries where they do not know the answer, even with access to an external knowledge base.
- **Group-conditional Conformal Factuality** Conformal methods that provide group-conditional factuality guarantees rather than simple marginal factuality guarantees.
- **Distribution-shift Aware Conformal Factuality** Conformal methods that provide factuality guarantees robust to different types of distribution shift, such as covariate shift.
- **Abstention-aware Conformal Factuality** Conformal methods for guaranteeing the factuality of LLM outputs by selectively abstaining at the sub-claim level.

Using this taxonomy, we classify 106 conformal methods that meet the selection criteria described in Section 2(a). The classification is performed on the 2^{nd} level (see Figure 2).Table 2 surveys all 106 conformal methods, extracting structured metadata described below:

- **Language Model(s)** The language model(s) which are conformalised in the study. We stratify language models into the following categories based on their parameter counts: language model (LM) if < 1 B, SLM if < 6 B, or LLM if > 6 B. For dense language models, we show their total parameter counts in brackets if they are not part of the model name, and for Mixture of Experts (MoE) language models, we show both their total parameters and their activated parameters in brackets.
- **Task(s)** The task(s) that are the primary objective(s) of the study, where conformal method(s) are proposed to ensure the validity and efficiency of language model(s) in achieving the task(s). An example includes multiple-choice question-answering (MCQA).
- **Conformal Method(s)** Conformal method(s) are given as the algorithm(s) proposed in each study, followed by the conformal framework(s) the algorithm(s) are based on. For details on several frameworks and their variants, refer to the preliminary (Section 4).
- **Dataset(s)** The dataset(s) that the conformal method is evaluated on; usually task-based.
- **Evaluation Metric(s)** The metric(s) reported in the study or other studies that evaluate the conformal method. A list of metrics and their definitions is provided in Table 2.

The remaining sections are each dedicated to a specific branch node in Figure 2, with the exception of the open-ended and close-ended conformal calibration branch nodes, which are combined for ease of reading. Subsections are stratified according to associated leaf nodes. The sections are organised as follows. Section 6 presents patterns and trends in research on conformal methods applied to LLMs. Section 7 surveys split-conformal-based methods, Section 8 selective conformal methods, Section 9 abstention-aware conformal methods, Section 10 conformal factuality methods, Section 11 conformal methods for retrieval-augmented generation, and Section 12 conformal risk control-based methods. Sections 7, 8, 11 include performance tables with the following metadata: conformal method(s), evaluation metric(s), and performance notes. A set of definitions for the evaluation metrics and their complements across the performance tables is provided in Table 2. Where metrics are specific to a particular study or group of studies, we encourage the reader to refer to the details in those papers.

Table 1: We present detailed characteristics of all 106 conformal prediction for LLM studies we identified in our extensive literature search. The following applies to language models: ** denotes a language model, * denotes a small language model, and the remainder are large language models. The following applies to datasets: ◊ indicates the dataset is unique to the study and not an established benchmark, and ◊ indicates the method is evaluated on the dataset in another study. The following applies to metrics: ⋆ indicates the metric is not reported and is computed from reported metrics, ⋆ indicates the metric is reported for the method in another study, and ⋆ indicates that the metric is reported initially but we take the same metric reported for the method in another study instead. Conformal methods are colour-coded based on their classification in the taxonomy illustrated in Figure 2; close-ended calibration, open-ended calibration, both open and closed-ended calibration, conformal RAG, conformal selective prediction, conformal factuality, and conformal abstention.

Ref	Language Model(s)	Task(s)	Conformal Method(s)	Dataset(s)	Evaluation Metric(s)
[50]	LLaMA-13B	MCQA	SCP with LAC	MMLU	ECR, APSS, SSC, ECE, MCE
[51]	OPT-125M**, OPT-350M**, OPT-1.3B*, OPT-2.7B*, OPT-6.7B, OPT-13B, OPT-30B	Text generation	CNS: Conformal nucleus sampling (based on SCP)	English Wikipedia Dump, OpenWebText	ECR⋆, APSS⋆, SSC⋆, calibrated probability-mass threshold
[52]	PaLM-2L, PaLM-2L-IF, GPT-3.5	RTP (as MCQA)	KnowNo: Single and multi-step uncertainty alignment (based on SCP)		
[53]	ViT+GPT-2-small (124M)**, T5-XL (3B)*, LLaMA-13B	RRG, text summarisation, OpenQA, reasoning, math	CLM: Conformal language modelling (based on conformal sampling with rejection using LTT with PAC)	MIMIC-CXR, CNN/DM, TriviaQA, GSM8K◊	mean prediction set loss, normalised set size, relative excess samples, area under normalised set size-risk curve, ECR⋆, APSS⋆, SSC⋆
[54]	CodeLLaMA-7B, FLAN-T5-XXL-11.3B, Falcon-40B-Instruct	Code generation, text summarisation, LLM-alignment	PRC: Prompt risk control (based on LTT, QRC [33], and statistical dispersion control [49])	MBPP, Helpfulness and Harmlessness, Red Teaming, MeQSum	Mean pass@10 loss, median toxicity VaR, Gini coefficient of loss
[55]	Mistral-7B-Instruct-v0.1	Stepwise decision planning (as MCQA)	Point-wise dependency neural estimation with conformal thresholding (based on SCP)	Synthetic Smart-home Decision Planning◊	PPV, TPR, F-measure
[56]	M2M100-400M**, M2M100-1.2B*, OPT-350M**, OPT-1.3B*	Text generation, machine translation	NE-CNS: Non-exchangeable conformal nucleus sampling (based on NECP [43] with nearest neighbours)	English Wikipedia Dump, OpenWebText, WMT-2022	ECR, APSS, SSC, expected coverage gap, calibrated probability mass threshold
[57]	T5 (220M)**, T5-small (60M)**	Reasoning, math, text generation	CBS: Conformal beam subsets (based on group-conditional SCP [40]) + DCBS: Dynamic conformal beam search (based on SCP with stepwise thresholds)	Integer Addition◊, USPTO-MIT	Beam coverage, conditional coverage, global coverage, MAE, ECR, APSS
[58]	N/A	RAG, OpenQA	CONFLARE: Conformal large language model retrieval (based on SCP)	N/A	N/A
[59]	Gemini Pro	OpenQA	Conformal abstention (based on CRC and RCPS [37])	Temporal Sequences, TriviaQA	Empirical hallucination risk, AR
[60]	GPT-3.5-Turbo, LLaMA-2-7B	RAG, OpenQA	TRAQ: Trustworthy retrieval augmented question answering (based on SCP with Bayesian optimisation) + TRAQ-P (TRAQ with PAC)	Natural Questions, TriviaQA, SQuAD-1, BioASQ	ECR, APSS
[61]	LLaMA-2-7B, LLaMA-2-13B, Mistral-7B-Instruct-v0.2	RAG, text summarisation, text generation	C-RAG: Conformal generation risk (based on CRC)	AESLC, CommonGen, DART, E2E	Conformal generation risk, empirical generation risk, Hellinger distance ρ

Continued on next page

Table 1 continued from previous page

Ref	Language Model(s)	Task(s)	Conformal Method(s)	Dataset(s)	Evaluation Metric(s)
[62]	GPT-4V	EQA (as MCQA)	Semantic exploration planner with calibrated stopping rule (based on SCP)	HM-EQA	Success rate, efficiency
[63]	LLaMA-2-7B-Chat	MCQA, sentiment analysis	IVAP: Inductive Venn-Abers predictor [64] for LLM calibration	BoolQ, Stanford Sentiment Treebank	ECE, AUROC, Brier loss, F-measure
[65]	GPT-4	OpenQA, reasoning, math, biography generation, fact verification	Conformal factuality (based on nested SCP with entailment sets)	FActScore, Natural Questions, MATH	EF, CRR−, CRR+*
[66]	LLaMA-2-13B, Mistral-7B-Instruct	OpenQA, reading comprehension	SCP with clustering by semantic equivalence	TriviaQA, CoQA	Accuracy (≈ECR), APSS, AUROC, AUARC, AURAC
[67]	LLaMA-3.1-70B-Instruct, GPT-4, GPT-4o	Reasoning, math	Coherent factuality (based on SCP with deducibility graphs and subgraph filtering)	MATH, FELM	EF, CRR+, correct final-answer rate, reprompting error rate reduction, legibility error detection rates
[68]	LLaMA-3-8B, Prismatic VLM	EQA, IQA, reasoning, OpenQA	Conformal reasoning: Offline adaptation + Refresh prediction + Guided reasoning (based on ACI [48])	20 Questions, Auto20Q, MediQ, HM-EQA	ECR, percent answered, TNR, efficiency, APSS*, NSR*, AR*
[69]	VideoLLaMA-7B, VideoLLaMA-13B, PandaGPT-7B, PandaGPT-13B, NExT-GPT (7B), Gemini-1.5-Flash, Gemini-1.5-Pro, GPT-4o-mini, GPT-4o, LLaMA-3-8B-Instruct, LLaMA-3.1-8B-Instruct, LLaMA-3-70B-Instruct, LLaMA-3.1-70B-Instruct	OpenVideoQA, MCVideoQA, OpenQA, VQA, MCQA, reading comprehension	TRON: Conformal score-guided sampling and nonconformity-based identification (based on SCP)	Video-MME, NExT-QA, MUSIC-AVQA, MSVD, TriviaQA, CoQA, VQA, MMLU	EER, ECR*, accuracy, APSS, SSM, SSC*
[70,71]	LLaMA-3-8B	KGQA	UAG: Calibrated knowledge graph traversal (based on SCP) + Error-rate control (based on LTT)	WebQuestionSP, Complex WebQuestions	ECR, APSS
[72]	DeepSeek-Prover-V1.5-RL-7B, DeepSeek-Prover-V1-7B, DeepSeek-Prover-V1.5-7B, GPT-4	Formal theorem proving	Stratified conformal proof search (based on SCP)	MiniF2F, ProofNet	Pass rate@passes, validation pass rate, test pass rate, number of passes required, wall clock time
[73]	LLaMA-2-7B, LLaMA-2-13B, WizardLM-13B-v1.2, Vicuna-7B-v1.5	MCQA, OpenQA	LofreeCP: Logit-free conformal prediction (based on SCP)	MMLU, TriviaQA, WebQuestions	ECR, APSS, SSC
[73,74]	LLaMA-2-7B, LLaMA-2-13B, WizardLM-13B-v1.2, Vicuna-7B-v1.5	OpenQA	SAPS: Sorted adaptive prediction sets	TriviaQA, WebQuestions	ECR*, SSC*, APSS*
[75]	GPT-3.5-Turbo, LLaMA-2-7B-Chat, Mistral-7B-Instruct-v0.3, LLaMA-3-8B-Instruct, Vicuna-13B-v1.5, LLaMA-2-13B-Chat, LLaMA-3-70B-Instruct	OpenQA, reading comprehension, MCQA	ConU: Conformal uncertainty (based on SCP)	CoQA, TriviaQA, MedQA, MedMCQA	AUROC, ECR, APSS, model accuracy
[76]	ViT+GPT-2-small (124M)**	RRG	OCS: Optimised conformal selection	MIMIC-CXR	EFDR, power
[77]	BGE-large-1.5 (326M)**, E5-Mistral-7B	RAG	SCP with score refinement	FEVER, SCIFACT, FIQA	ECR, APSS

Continued on next page

Table 1 continued from previous page

Ref	Language Model(s)	Task(s)	Conformal Method(s)	Dataset(s)	Evaluation Metric(s)
[78]	LLaMA-3-8B-Instruct, Gemini-Pro, GPT-3.5-Turbo	RAG, DAO, event detection, event argument extraction	DRAG: Diverse-RAG retrieval + AdaCP: Adaptive conformal prediction (based on SCP with threshold decay)	ACE05-E, CASIE	Event detection, event argument extraction, combined event extraction
[79]	GPT-3.5-Turbo, GPT-4o	OpenQA, biography generation, fact verification	CB: Conditional boosting + LA: Level-adaptive conformal prediction (based on QCCP [38])	MedLFQA, Wikipedia Biographies, FActScore	EF, CRR+, EMR, GCEF
[80]	LLaMA-2-7B, LLaMA-2-13B, LLaMA-2-70B, Mistral-7B, Mixtral-8x7B, Falcon-7B, Falcon-40B, MPT-7B, Gemma-7B, Qwen-1.8B*, Qwen-7B, Qwen-14B, Qwen-72B, Yi-6B, Yi-34B, DeepSeek-7B, DeepSeek-67B, InternLM-7B, GPT-3.5, GPT-4	MCQA, reading comprehension, commonsense inference, DRS, document summarisation, OpenQA	SCP with LAC and APS	MMLU, CosmosQA, HellaSwag, HaluDial (from HaluEval), HaluSum (from HaluEval), TriviaQA	Accuracy, ECR, APSS, ECE
[81]	ViT+GPT-2-small (124M)**, OPT-13B, LLaMA-2-13B-Chat	RRG, OpenQA, reading comprehension	Conformal alignment (based on conformalised selection)	MIMIC-CXR, TriviaQA, CoQA	EFDR, power
[82]	LLaMA-2	MCQA	CPL: Split conformal prediction with length-optimisation	TruthfulQA, MMLU, OpenBookQA, PIQA, BigBench	ECR, APSS, NSR*
[83]	GPT-3.5, GPT-4-Turbo	RTP (as MCQA), RAG	Introspective conformal prediction (based on SCP)	Mobile Manipulation◇, Safe Mobile Manipulation◇, Tabletop Rearrangement◇	Success rate (≈ECR), help rate (≈NSR), exact set rate, non-compliant contamination rate, unsafe contamination rate, overask rate, overstep rate, unsafe rate, APSS
[84]	GPT-3.5-Turbo, Alpaca-7B	OpenQA	SGenSemi: Semi-supervised selective generation with entailment pseudo-labelling (based on PAC CP [47])	Natural Questions, QA2D	Empirical FDR by textual entailment (FDR-E), efficiency
[85]	GPT-3.5, LLaMA-2-7B, LLaMA-3-8B	RTP (as MCQA)	S-ATLAS: Safe planning for teams of language-instructed agents (based on SCP)	AI2THOR Simulated Home-service Scenarios◇	Mission success rate (≈ECR), help rate (≈NSR), number LLM queries required
[86]	LLaMA-3.2-1B*, Qwen-2.5-3B-Instruct*, Vicuna-7B-v1.5, Qwen-2.5-7B-Instruct, LLaMA-3.1-8B-Instruct, Vicuna-13B-v1.5	OpenQA, reading comprehension	TECP: Token-entropy conformal prediction (based on SCP)	TriviaQA, CoQA	EMR, ECR*, APSS
[87]	Mistral-7B-Instruct-v0.2, GPT-3.5-Turbo, GPT-4-Turbo	Text summarisation preference, chat assistance preference, response preference, LLM-as-a-judge	SA: Simulated annotators + CSE: Cascading selective evaluation (based on CRC)	AlpacaEval, TL;DR, ChatArena, Auto-J	Empirical human agreement, empirical disagreement risk*, coverage rate, AR*, guarantee success rate, accuracy, ECE, AUROC, AUPRC, pipeline accuracy*

Continued on next page

Table 1 continued from previous page

Ref	Language Model(s)	Task(s)	Conformal Method(s)	Dataset(s)	Evaluation Metric(s)
[88,89]	Yi-34B, Qwen-7B, Qwen-14B, LLaMA-2-7B, LLaMA-2-13B, LLaVA-v1.6-7B, LLaVA-v1.6-13B, LLaVA-v1.6-34B, LLaVA-Phi2-2.7B*, Monkey-Chat-7B, InternLM-XComposer2-VL-7B, Yi-VL-6B, CogAgent-VQA-7B, MobileVLMV2-7B, Mplug-Owl2-7B, Qwen-VL-Chat-7B	MCQA, reading comprehension, commonsense inference, DRS, document summarisation, MCVQA, OpenQA, math, reasoning, diagram understanding, fact verification	CAP: Conformalised abstention policy (based on SCP with reinforcement learning)	MMLU, CosmosQA, HellaSwag, HaluDial (from HaluEval), HaluSum (from HaluEval), MMBench, Digits (subset of OODCV-VQA), ScienceQA, SEEDBench, AI2D, FActScore, MATH, Natural Questions	ECE, accuracy, ECR, APSS, AUROC, AUARC, CRR+
[90]	GPT-Neo-1.3B*	Text classification	ConfTS: Conformal temperature scaling for SCP (with APS and RAPS)	AG News, Dbpedia	ECR, APSS
[91]	LLaVA-OneVision-Qwen2-7B, LLaVA-1.5-13B, LLaVA-OneVision-Qwen2-7B-VisCoT, Vicuna-7B-VisCoT, SPHINX-13B	VQA	SRICE: Split quasi-conformal uncertainty for pathway selection + Conformal tool calibration	VQA2, VizWiz, Flickr30K, GQA, MMBench	Accuracy, ECE
[92]	LLaMA-3.1-8B-Instruct	Reasoning, NL2ASP	CLM: Conformal language modelling (from [53])	StepGame	Task accuracy
[93]	GPT-4o	LLM-as-a-classifier	Conformal tree: tree-based local conformal calibration (based on SCP)	UCI Erythemato-Squamous Disease, Voteview DW-NOMINATE Legislator Ideology$^{\diamond}$	Class-wise prediction accuracy, ECR, APSS
[94]	LLaVA-1.5-7B, LLaVA-1.5-13B, LLaVA-v1.6-Mistral-7B, LLaVA-v1.6-Vicuna-13B, Qwen-2-VL-2B-Instruct**, Qwen-2-VL-7B-Instruct, InternVL2-1B**, InternVL2-8B	MCVQA	SCP with LAC	MMMU, ScienceQA	EER, ECR*, APSS
[95]	GPT-3.5-Turbo, GPT-4o, Falcon-7B, LLaMA-3.1-7B-Instruct	AVP (as MCQA)	SafePath: Path generation, uncertainty-aware path selection, path decision (based on SCP)	nuScenes, highway-env	Deviation from target coverage, ECR*, APSS, human delegation rate
[96]	N/A	RTP (as OpenQA), NL2LTL	ConformalNL2LTL: Natural language to linear temporal logic (based on SCP)	Synthetic NL2LTL Tasks$^{\diamond}$	Translation accuracy (≈ECR), help rate (≈NSR), fraction of formulas needing help
[97]	LLaMA-3.2-1B-Instruct*, LLaMA-3.2-3B-Instruct*, Qwen-2.5-1.5B-Instruct*, Qwen-2.5-3B-Instruct*	MCQA	SCP with self-consistency theory	MedMCQA, MedQA, MMLU	EMR, ECR*, APSS, AUROC
[98]	ViT+GPT-2-small (124M)**, T5-XL (3B)*, LLaMA-7B	RRG, text summarisation, OpenQA, reasoning, math	SCOPE-Gen: Sequential conformal prediction for generative models (based on SCP with Markov factorisation)	MIMIC-CXR, CNN/DM, TriviaQA, BBH Date Understanding$^{\diamond}$	Admission query rate, calibration runtime, ECR*, APSS, calibration rejection rate (≈AR)
[99]	LLaMA-3-8B, Qwen-2.5-14B, DeepSeek-V3 (671B, 37B), DeepSeek-R1-Distilled-LLaMA-3-8B, DeepSeek-R1-Distill-Qwen-14B, DeepSeek-R1 (671B, 37B)	LLM-routing, MCQA, math, reasoning, reading comprehension	CP-Router: SCP with full and binary entropy-based adaptive error-rate calibration	MMLU-STEM, STEM-MCQA, GSM8K, GPQA, LogiQA, CN-Chemistry	Accuracy, token reduction ratio, token utility

Continued on next page

Table 1 continued from previous page

Ref	Language Model(s)	Task(s)	Conformal Method(s)	Dataset(s)	Evaluation Metric(s)
[100]	LLaMA-3-8B-Instruct, Phi-3-mini-Instruct (3.8B)*, Gemma-2-9B-SimPO, GPT-4	MCQA, NL2SQL, tool selection	CROQ: Conformal revision of questions + CP-OPT: Conformal prediction optimisation (based on SCP)	MMLU, TruthfulQA, ToolAlpaca, BIRD	Accuracy, ECR, APSS
[101]	Phi-2 (2.7B)*, LLaMA-3-8B, GPT-4o-mini, Gemma-2-27B	OpenQA, code generation, reasoning, math	GPS: Generative prediction sets (based on CQR [44])	MBP-P, DS-1000, GSM8K, MATH, TriviaQA	ECR, APSS, AR, number samples generated, non-abstention empirical coverage
[102]	Qwen-2.5-3B-Instruct*, Qwen-2.5-7B-Instruct, LLaMA-3.1-8B-Instruct, Vicuna-13B-v1.5, Qwen-2.5-14B-Instruct, Qwen2-VL-2B-Instruct*, Qwen2-VL-7B-Instruct, InternVL2-8B, LLaVA-1.5-13B, LLaVA-v1.6-Vicuna-13B	OpenQA, commonsense inference, MCQA	COIN: Calibration set error analysis, upper bound construction, and threshold selection with PAC (based on LTT)	TriviaQA, CommonsenseQA, MMVet	Empirical conditional failure rate (≡EFDR), power
[103]	LLaMA-2-7B	Text generation, LLM-alignment	Conformal distortion risk control (based on CRC and L-statistics)	RealToxicityPrompts	VaR, CVaR, average sampling cost
[104]	BLIP-2+FLAN-T5-XL (3B)*	SGG (as MCQA)	PC-SGG: Plausibility-ensured conformal scene graph generation (based on Mondrian SCP)	Visual Genome (VG150)	ECR, APSS, average class coverage gap, triplet coverage validity, coverage-Recall@K, coverage-mean-Recall@K
[105]	LLaMA-2-7B, LLaMA-3-8B, Mistral-7B	LLM-recommendation	FACTER: Fairness aware conformal thresholding and prompt engineering (based on SCP)	MovieLens-1M, Amazon Movies & TV	Sub-network similarity ratio, sub-network similarity variance, counterfactual fairness ratio, violation rate, type-I error, detection power, NDCG@10, recall@10
[106]	LLaMA-3-8B-Instruct, Mixtral-8x7B-Instruct	Reasoning, math	CPQ: Conformal prediction with query oracle (based on SCP with missing mass optimisation)	BBH Geometric Shapes, GSM8K, BBH Date Understanding	ECR, APSS, EE fraction
[107]	LLaMA-3.1-8B, Qwen-2.5-7B, Phi-3-small (7B)	IQA, MCQA, reasoning	C-IP: Conformal information pursuit (based on SCP with predictive inference)	20 Questions (subset of AwA2)*, MediQ	ECR, accuracy
[108]	LLaMA-3.2-3B-Instruct*, Qwen-2.5-3B-Instruct*, Qwen-2-7B-Instruct, Qwen-2.5-7B-Instruct, OpenChat-3.5 (7B), LLaMA-3-8B-Instruct, LLaMA-3.1-8B-Instruct, LLaMA-2-13B-Chat, Vicuna-13B-v1.5, Qwen-2.5-14B-Instruct, Qwen-2.5-32B-Instruct	OpenQA, MCQA, reasoning, reading comprehension	SConU: Selective conformal uncertainty (based on SCP)	MMLU, MMLU-Pro, MedMCQA, TriviaQA, CoQA	EMR, ECR*, SSM, APSS, SSC*
[109]	ViT+GPT-2-small (124M)**, OPT-13B, LLaMA-2-13B-Chat	RRG, OpenQA, reading comprehension	ACS: Adaptive conformal selection	MIMIC-CXR, TriviaQA, CoQA	EFDR, power, expected similarity

Continued on next page

Table 1 continued from previous page

Ref	Language Model(s)	Task(s)	Conformal Method(s)	Dataset(s)	Evaluation Metric(s)
[110]	LLaMA-2-7B-Chat, Mistral-7B-Instruct-v0.2	Biography generation, fact verification	MVSC: Multivalid SCP + GCCQR: Group-conditional CQR [44]	Bio-NQ (subset of Natural Questions)*, Bio-FActScore	EF, mean absolute groupwise coverage error, biography retention rate, biography facts retention rate, average squared calibration error (ASCE), max/mean group ASCE (gASCE), Brier score, CRR+*, AR*, GCEF intervals*
[111]	GPT-4o	RAG, OpenQA, reasoning, fact verification	Conformal-RAG: Group-conditional conformal factuality for RAG with context similarity scores (based on [65] and Mondrian SCP [41])	MedLFQA, FActScore, PopQA, HotpotQA	EF, CRR−, GCEF, CRR+*
[112]	Gemini-1.5-Flash	RAG, fact generation	RAG-based conformal factuality with reference model confidence scoring (based on [65])	FActScore	EF, power (≡claim TPR), claim FPR, CRR+*
[113]	ChatRhino-4B*, ChatRhino-10B, ChatRhino-81B	Geospatial repartition	CoAlign: Conformal ranking for human-aligned LLM geospatial repartition with transductive adjustment	Offline Logistics Repartition$^\diamond$, Online Logistics Repartition A/B$^\diamond$	ECR, prediction set ratio, false coverage rate (≈EMR), human intervention rate, recommendation acceptance rate, method similarity ratio, maximum similarity, workload distribution balance
[114]	Qwen-2.5-3B-Instruct*, LLaMA-3.2-3B-Instruct*, Qwen-2.5-7B-Instruct, Meta-LLaMA-3-8B-Instruct, LLaMA-3.1-8B-Instruct, Vicuna-7B-v1.5, OpenChat-3.5 (7B)	MCQA, reasoning	SCP with significance testing using conformal p-values	MMLU, MMLU-Pro	EER, ECR*, APSS
[115]	LLaMA-2-13B-Chat	OpenQA	OCS-ARC: Online conformal selection with accept-to-reject changes	TriviaQA	FDP (≈EFDR), power
[116]	Qwen-VL-2.5-7B-Instruct	VQA	NED-CP: Normalised edit-distance conformal calibration for abstention (based on SCP)	OK-VQA	Selective prediction coverage, accuracy, exact match accuracy, normalised edit distance, AR*
[117]	DeepSeek-V3 (671B, 37B), GPT-4o, GPT-4o-mini, Kimi-K2 (1T, 32B), DeepSeek-Chat (671B, 37B)	RAG, MCQA	ConfAgents: Adaptive multi-agent framework with a conformal-gated triage step (based on SCP)	MedQA, MMLU, MedBullets, AfriMedQA	Accuracy, processing time, completion tokens, total tokens, balanced efficiency score, preference rate
[118, 119]	Vicuna-7B-v1.5, Vicuna-13B-v1.5, Qwen-2.5-3B-Instruct*, Qwen-2.5-7B-Instruct, LLaMA-3.2-1B*, LLaMA-3.1-8B-Instruct	MCQA, reasoning	SCP with frequency-based predictive entropy	MedMCQA, MedQA, MMLU, MMLU-Pro	EMR, ECR*, APSS, AUROC
[120]	GPT-3.5, LLaMA-2-13B, LLaMA-3-8B, Qwen-3-32B	RTP (as OpenQA), NL2LTL	HERACLES: Hierarchical conformal natural language planner (based on SCP)	Synthetic NL2LTL Tasks$^\diamond$, Mini-City Hardware Scenarios$^\diamond$	Mission success rate (≈ECR), help rate (≈NSR), average length of successful plans

Continued on next page

Table 1 continued from previous page

Ref	Language Model(s)	Task(s)	Conformal Method(s)	Dataset(s)	Evaluation Metric(s)
[121]	Gemini-2.0-Flash, GPT-3.5-Turbo	MCQA, math, reasoning, LLM-as-a-judge	Inter-cascade: Interactive online LLM cascade with in-context knowledge distillation with PAC (based on LTT)	GLM-Symbolic, GSM-Plus, MetaMath, NASA-History-MCQ	Pipeline accuracy, weak LLM accuracy, coverage rate, AR*, strong LLM call rate, weak correct accepted, token reduction, cost reduction
[122]	N/A	OpenQA, code generation, RAG, MCQA	UniCR: Unified confidence calibration and refusal (based on CRC)	RAG LFQA Dataset$^\diamond$, TruthfulQA, SFQA Prompts$^\diamond$, Code Generation Benchmark$^\diamond$	Risk-coverage curves, area under the risk-coverage curve, empirical selective risk, test coverage, violation frequency, FActScore, exact-match rate, contradiction rate, end-to-end latency, AR*
[123]	GPT-4.1-nano, Gemini-2.5-Pro, Gemini-2.5-Flash	Text classification, topic prediction, sentiment analysis	CondCP-Filter: Conformal filtering via kernel-localised QCCP [38]	Symptom to Diagnosis, arXiv Abstracts$^\diamond$, Emotions Twitter	F-measure, PPV, TPR, Shannon entropy
[124]	LLaVA-v1.6-13B, Monkey-Chat (7B), LLaVA-v1.6-7B, InternLM-XComposer2-VL (7B), Yi-VL-6B, CogAgent-VQA (7B), MobileVLMV2-7B, MoE-LLaVA-Phi2-2.7B*, mPLUG-Owl2 (7B), Qwen-VL-Chat (7B), QLaVA-v1.6-34B	MCVQA, diagram understanding	SCP (using LAC and APS)	MMBench, Digits (subset of OODCV-VQA), ScienceQA, SEEDBench, AI2D	ECR, accuracy, uncertainty-aware accuracy, APSS, ECE, MCE
[125]	N/A	Text generation	CoVeR: Token-level conformal calibration for autoregressive decoding (based on PAC CP [47])	N/A	N/A
[126]	GPT-4.1-nano, GPT-4o-mini, LLaMA-4-Scout (109B, 17B), Gemma-3-4B*, Gemma-3-12B, Gemma-3-27B, InternVL3-1B*, InternVL3-2B*, InternVL3-8B, Qwen-2.5-VL-3B*, Qwen-2.5-VL-72B, LLaVA-1.5-7B, LLaVA-1.5-13B, Molmo-E-1B*, Molmo-D-7B, Pixtral-12B	MCVQA, reasoning, diagram understanding	Split conformal prediction (using LAC, APS, marginal score)	MMMU, MMMU-Pro, ScienceQA, AI2D, MathVision, WorldMedQAV	ECR, APSS, accuracy, entropy score
[127]	LLaMA-3.3-70B-Instruct-Turbo, GPT-4o-mini, Mixtral-8x7B-Instruct-v0.3, LLaMA-3.1-8B-Instruct-Turbo	OpenQA, reasoning	B-UCP: Batched unsupervised CP + BB-UCP: Batched bootstrap UCP + Batch-level predicate conformal alignment	ASQA, NQ-Open, HotpotQA, AmbigQA	ECR, average interval width (≈APSS), factuality severity, factuality lift, average coverage gap
[128]	LLaMA-3.3-70B, Mixtral-8x7B, LLaMA-3.1-8B, GPT-4o-mini, Meta-LLaMA-3.1-8B-Instruct-Turbo	OpenQA, LLM-as-a-judge, reasoning	BB-CRC: Batched bootstrap CRC + RBWA-CRC: Randomised batch weighted-average CRC + Conformal actuator monotone decision gate	ASQA, NQ-Open, HotpotQA, AmbigQA	Empirical risk, factuality severity, factuality severity reduction, LLM-as-Judge severity

Continued on next page

Table 1 continued from previous page

Ref	Language Model(s)	Task(s)	Conformal Method(s)	Dataset(s)	Evaluation Metric(s)
[129]	LLaMA-3-8B	OpenQA, math, reasoning	Similarity-weighted CRC for hallucination detection	TriviaQA, GSM8K	Hallucination detection TPR, TPR variability under shift, hallucination miss risk (≡FNR)*
[130]	Moonshot-v1-8k, LLaMA-2-7B, LLaMA-2-13B, LLaMA-3-8B-Instruct, LLaMA-3-70B-Instruct, Qwen-Plus, Qwen-32B, DeepSeek-V3 (671B, 37B), DeepSeek-R1 (671B, 37B), GPT-3.5-Turbo, GPT-4o	OpenQA	AggLCF: Aggregation-enhanced localised conformal factuality (based on SCP with engression and cluster-level conformity)	MedLFQA	EF*, EMR, number retained subclaims, CRR+*
[131]	GPT-4.1-nano, GPT-4.1-mini, GPT-4.1	MCQA, LLM-alignment	Conformal arbitrage (based on CRC)	TruthfulQA, MMLU, PKU-SafeRLHF	Accuracy, average cost per example, empirical human alignment, safety-violation loss, accuracy-loss guardrail risk*
[132]	LLaMA-2-7B, LLaMA-2-13B, LLaMA-2-70B, Mistral-7B, Falcon-7B, Falcon-40B, MPT-7B, Qwen-1.8B*, Qwen-7B, Qwen-14B, Qwen-72B, Yi-6B, Yi-34B, DeepSeek-7B, DeepSeek-67B, InternLM-7B	MCQA	DS-CP: Domain-shift-aware conformal prediction (based on weighted SCP [42])	MMLU	ECR, APSS
[133]	LLaMA-3.1-8B-Instruct, OpenChat-3.5 (7B), Qwen-2.5-3B-Instruct*, Qwen-2.5-7B-Instruct, Qwen-2.5-14B-Instruct	OpenQA, MCQA, reading comprehension	SAFER: Abstention-aware sampling (based on LTT) + Conformalised filtering (based on CRC)	TriviaQA, ScienceQA, CoQA	EER, ECR*, APSS, average calibrated sampling budget
[134]	LLaMA-3-8B-Instruct, LLaMA-3-7B-Instruct, GPT-4o, Qwen-2.5-72B-Instruct	OpenQA, biography generation, fact verification	CLC: Conformal linguistic calibration for uncertainty-aware claim rewriting (based on LTT)	SimpleQA, Natural Questions, FActScore	EF, conditional pointwise mutual information
[135]	Mistral-7B-Instruct-v0.3, Gemma-7B-it, LLaMA-3.1-8B-Instruct	OpenQA, MCQA, reasoning	AR-NECP: Label-conditional adaptive rejection and non-exchangeable CP (based on SCP)	TriviaQA, HotpotQA, MMLU	ECR, APSS, unanswerable APSS, mean relative error of estimated domain counts
[136]	Qwen-2.5-Math-7B-PRM800K, Math-Shepherd-PRM-7B, UltraRM-13B, GPT-4	Math, reasoning, property-constraint satisfaction	E-scores: e-value calibration for post-hoc correctness guarantees of generative outputs (based on conformal e-prediction [45])	ProcessBench, UltraFeedback	Size distortion, mean error vs. mean tolerance level, precision-recall curves, AUPRC*
[137]	GPT-3.5-Turbo, GPT-4o, GPT-4o-mini	OpenQA, biography generation	CoFact: Online density-ratio estimation + Adaptive reweighted SCP for covariate shift	MedLFQA, Wikipedia Biographies, WildChat+ (from WildChat)	EF, CRR+
[138]	DeepSeek-R1-Distill-Qwen-1.5B*, DeepSeek-R1-Distill-LLaMA-70B, LLaMA-3.1-8B-Instruct, Qwen-2.5-7B-Instruct, Qwen-2.5-32B-Instruct, QwQ-32B, s1.1-7B, s1.1-32B, Skywork-OR1-7B, Skywork-OR1-32B	Math, reasoning	ATTS: Asynchronous test-time scaling with online-calibrated conformal rejection sampling	MATH100, AIME24, AIME25, AMC23	Marginal accuracy, conditional accuracy, budget prediction accuracy, end-to-end speedup, latency, throughput, token consumption
[139]	N/A	OpenQA, RRG	MCCS: Multi-condition conformal selection	N/A	N/A

Continued on next page

Table 1 continued from previous page

Ref	Language Model(s)	Task(s)	Conformal Method(s)	Dataset(s)	Evaluation Metric(s)
[140]	LLaMA-3.1-8B, Mistral-7B-Instruct-v0.3, DeepSeek-R1-Distill-Qwen-14B	MCQA, sentiment analysis, fact verification, topic prediction, toxicity detection, LLM-alignment, OpenQA, reading comprehension, commonsense inference	Conformal probes: Confidence-threshold selective prediction with deferral (based on SCP)	MMLU, CosmoQA, PiQA, ARC, MedMCQA, CommonsenseQA, OpenBookQA, QASC, AmazonReviews, TwitterSentiment, Yelp Polarity, TwitterFinance, NewsMTC, IMDB Reviews, Financial Phrasebank, AuditorSentiment, DAIR-AI Emotion, SST5, ClimateFEVER, HealthVER, FEVER, AGNews, BBCNews, NYTimes, JigsawToxicity, JigsawUnintendedBias-Toxicity, Self-Aware, KnownUnknown, WildJailbreak, NaturalQA, MCMarco, TriviaQA	Consistency (≈non-deferred subset accuracy), selective prediction coverage, PPV, TPR, accuracy loss, AR*
[141]	LLaMA-3.1-8B, Phi-3-small (7B)	MCQA, reading comprehension, commonsense inference, DRS, document summarisation	PA-QCCP: Paraphrase-aware nonconformity + SCP/QCCP [38] calibration	MMLU, CosmoQA, HellaSwag, HaluDial (from HaluEval), HaluSum (from HaluEval), MedMCQA, MedQA	ECR, APSS, accuracy, Brier score, NLL
[142]	Qwen-3-1.7B*, Qwen-3-4B*, Qwen-3-8B, Qwen-3-14B	LLM-routing, MCQA, reasoning, math, code generation	CR^2: Conformal risk-controlled routing with supervised contrastive learning answerability and multi-label candidate set	MMLU, MMLU-Pro, GSM8K, BigBench Hard, GPQA, MBPP	Routing accuracy, average per-sample token cost, average per-query composite routing loss, candidate set size (≈APSS)
[143]	Qwen-3-0.6B**, Qwen-3-0.6B-Think**, Qwen-3-4B*, Qwen-3-4B-Think*, Qwen-3-8B, Qwen-3-8B-Think, Qwen-3-32B, Qwen-3-32B-Think, Qwen-3-30B-A3B (30.5B, 3.3B), LLaMA-3.2-1B-Instruct*, LLaMA-3.2-3B-Instruct*, Meta-LLaMA-3.1-8B-Instruct, SmolLM2-123M-Instruct**, SmolLM2-360M-Instruct**, SmolLM2-1.7B-Instruct*, GPT-OSS-20B (21B, 3.6B), GPT-5-nano, GPT-5.1, Gemini-2.5-Pro	OpenQA, reasoning, math, RAG, fact verification	Split conformal factuality filtering for RAG (based on [65])	FActScore, FActScore-Rare, MATH, Natural Questions	EF, power (≡TPR), FPR, non-empty rate, non-vacuous empirical factuality, sufficient correctness, CRR+
[144]	N/A	OpenQA, RRG	MCS: Multivariate conformal selection	N/A	N/A
[145]	LLaMA-2-7B, LLaMA-3-8B, Mistral-7B, Falcon-7B, MPT-7B, Yi-6B	OpenQA, reading comprehension	COPU: Conformal prediction for uncertainty quantification (based on SCP using LAC)	TriviaQA, WebQuestions, SQuAD, WikiQA	ECR, APSS
[146]	LLaMA-3.1-8B-Instruct	MCQA	SOCOP: Singleton-optimised conformal prediction (based on SCP)	MMLU	ECR, APSS, NSR

Continued on next page

Table 1 continued from previous page

Ref	Language Model(s)	Task(s)	Conformal Method(s)	Dataset(s)	Evaluation Metric(s)
[147]	GPT-4o mini, Gemini 2.0 Flash-Lite, Gemini 2.5 Flash, LLaMA-3-8B, Qwen3-8B	Document summarisation	CIS: Conformal importance summarisation (based on SCP)	ECTSum, CSDS, CNN/DM, SciTLDR-AIC, SciTLDR-Full, MTS-Dialog	AUPRC, ECR, TPR, proportion sentences removed
[148]	LLaMA-3.3-70B-Instruct, Qwen-2.5-72B-Instruct, DeepSeek-V3 (671B, 37B)	OpenQA, biography generation	MACI: Multi-LLM adaptive conformal inference + MACI-DRE: MACI with density ratio estimation (based on SCP and Mondrian SCP)	MedLFQA, Wikipedia Biographies, ExpertQA	EF, CRR+, GCEF, MSE, FPR, Jaccard distance
[149]	GPT-4o, GPT-4o-mini, DeepSeek-R1-Distill-Qwen-32B, Qwen-2.5-72B-Instruct	LLM-as-a-judge, reasoning, text summarisation, reading comprehension	SCP with ordinal boundary adjustment (using CQR [44], Asym CQR [150], CHR [151], LVD [152], Boosted CQR/LCP [153], R2CCP [154], OrdinalAPS [155], OrdinalRC [156])	SummEval, DialSumm, CosmoQA, DROP, e-SNLI, GSM8K	ECR, average interval width (≈APSS), MSE, MAE, Spearson's rank correlation coefficient, Kendall's rank correlation coefficient
[157]	LLaVA-1.5, Phi-3.5-Vision-Instruct (4.2B)*, LLaMA-3.2-11B-Vision, GPT-4o-mini, LLaVA-Med-7B, CvT2DistilGPT2 (81M)**, MAIRA-2-7B, LLaVA-v1.6	RRG, document understanding, scene understanding	ConfLVLM: Conformal claim-level factuality filtering for LVLM (based on SCP)	MSCOCO, MIMIC-CXR, SROIE	EF, CRR−, AR, claim-level TPR/FNR/F-measure, response-level error rate/average loss, CRR+*
[158]	MentaLLaMA-7B-Chat, EYE-LLaMA-7B	OpenQA, MCQA, text classification	ICAD: Inductive conformal anomaly detection with dropout-tolerant nonconformity	COVID-QA, MedMCQA, EyeQA, IMHI	False alarm rate (≈EMR), ECR*, AUROC
[159]	Gemma-3-1B*	RTP (as MCQA)	CoFineLLM: Conformal finetuning for LLM-based planners (based on SCP)	BabyAI-Text◇, Mini-Urban GridWorld◇, Real Robot Mini-Urban◇	ECR, APSS, help rate (≈NSR), verification rate
[160]	LLaMA-3.2-3B*	RAG, LLM-recommendation	SarRec: Statistically-guaranteed augmented retrieval for recommendations (based on CRC with RCPS [37])	LastFM, Steam, MovieLens-100K	TPR, PPV, mean reciprocal rank, NDCG, APSS
[161]	GPT-4.1, o4-mini, o3-mini, LLaMA-3, LLaMA-4	EHR entity extraction, fact verification	Selective EHR extraction via conformal thresholding (based on SCP)	RespondHealth PD EHR notes (10K visits)◇, VeriFact-BHC (MIMIC-III)	Acceptance-set coverage (≈ECR), rejection rate (≈AR), Brier score, ECE, FActScore, PPV, TPR, F-measure, status accuracy, rating accuracy, inter-clinician agreement
[162]	GPT-4, Qwen-2-VL-7B, MiMo-VL-7B-RL	Arrhythmia detection, sleep stage classification	ConMIL: Conformalised multiple instance learning (based on CRC with RCPS [37])	SleepEDF, PTB-XL	Accuracy, PPV, TPR, FNR, F-measure, singleton rate, NSR, AR, ECR, APSS, AUROC, AUPRC
[163]	Mistral-7B, MetaMath-Mistral-7B, Zephyr-7B-beta, Chinese-Mistral-7B, Dolphin-2.6-Mistral-7B, LLaMA-3-8B, Dolphin-2.9-LLaMA3-8B	LLM-routing, MCQA, math, reasoning	RACER: Risk-aware calibrated efficient routing for multi-LLM systems (based on CRC)	GSM8K, MMLU, CMMLU, ARC-Challenge	Empirical misrouting risk, APSS, downstream aggregated accuracy, inference overhead reduction
[164]	LLaMA-3-8B-Instruct, Qwen2-VL-7B-Instruct, Qwen2.5-Math-RM-72B	OpenQA, math, reasoning, image captioning	CFC: Conditional factuality control (based on QCCP [38]) + CFC-P (CFC with PAC)	GSM8K, TriviaQA, Flickr8k	ECR, APSS, group-stratified coverage
[165]	GPT-4o, LLaMA-3.3-70B-Instruct, Qwen3-Embedding-8B	RAG	Context-RAG: Post-retrieval filtering using Conformal-Embedding and Conformal-LLM scoring (based on SCP)	NeuCLIR, RAGTIME	ECR, context removal rate, factual quality via ARGUE F-measure

Table 2: Evaluation metrics for the SCP, conformal selection, and conformal factuality methods.

Metric	Definition	Ref
Empirical Coverage Rate	ECR; Fraction of test examples for which the true output is contained in the prediction set.	[12]
Empirical Miscoverage Rate	EMR; The complement of ECR, such that EMR $= 1 -$ ECR. Also sometimes referred to as EER.	[12]
Size-stratified Coverage Rate	SSC; Coverage computed within bins of prediction set size; often summarised as the minimum bin-wise coverage.	[31]
Size-stratified Miscoverage Rate	SSM; The complement of SSC, such that SSM $= 1 -$ SSC.	[31]
Average Prediction Set Size	APSS; Average cardinality of the prediction set across test examples. In classification, this is the average number of outputs returned per example.	[12,166]
Non-singleton Rate	NSR; Proportion of test examples whose prediction set contains more than one output.	[166]
Empirical False Discovery Rate	EFDR; Proportion of selected outputs that are false discoveries, i.e. selected but negative/not aligned.	[34,81]
Power	Proportion of truly positive/aligned examples that are successfully selected.	[34,81]
Empirical Factuality	EF; Observed fraction of retained outputs that are factual. Equivalent to ECR.	[137]
Group-conditional Empirical Factuality	GCEF; The group-wise variant of EF.	
Claim Retention Rate	CRR+; Proportion of original claims retained after filtering.	[137]
Claim Rejection Rate	CRR−; The complement of CRR+; CRR− $= 1 -$ CRR+.	

Whilst we endeavoured to control variables such as user-specified error rate, tolerance level, failure probability, etc., in some cases, this was not possible. In addition, when multiple datasets were used within the same task, we report the mean performance across datasets for each method. Lastly, even in the ideal outcome where datasets and variables are equivalent across methods, there are still discrepancies in the dataset splits, temperature values, calibration set size, as well as confounding factors due to LLM families, evaluation protocols, task heterogeneity, etc. As such, these performance tables should only be taken as a guide, rather than an empirical comparison.

In light of this, we conduct a performance analysis with a strict selection criterion, such that the empirical results of various conformal methods are directly comparable. We conducted frequency analysis to isolate the most common datasets for evaluation, and the most common tasks LLMs are calibrated for using conformal methods (see Section 6(b) for results). We then pool the data and make selections that ensure a collection of datasets for a diverse list of tasks. We then look for conformal methods which are evaluated both for these tasks and on these datasets and group them by branch node, excluding any which do not report the two primary metrics of each branch. We then further refine our method selection by ensuring the nominal target coverage $1 - \alpha \in \{0.70, 0.90, 0.95, 0.99\}$ or nominal FDR $\alpha_{\text{FDR}} = 0.25$ be fixed. After narrowing down our list of conformal methods to evaluate, we add additional analysis: logit-based vs. logit-free methods, methods for LLMs vs. LVLMs, methods with marginal coverage vs. group-conditional coverage, and methods operating under various distribution shifts.

The performance of each conformal method is reported after the method is introduced, and is illustrated in Figure 6 for split-conformal-based methods with LLMs, Figure 8 for split-conformal-based methods with LVLMs, Figure 13 for conformal selection-based methods, and Figure 16 for conformal factuality methods. In the case of both the logit and language model comparisons, these are reported in their own respective sections, Section 7(a)ii and Section 7(b)ii, and in Figure 11. The evaluation metrics for the split-conformal, conformalised selection, and conformal factuality methods are given above; the metrics for the logit and language model comparisons are defined below for the nominal coverage in a given setting c, the observed coverage achieved by the conformal method $\hat{c}$ (ECR, see Table 2), and the number of evaluated settings over which the metric is averaged n. For aMACE, λ is the weighting coefficient that controls how much undercoverage vs. overcoverage is penalised (denoted λ_u and λ_o). APSS is defined in Table 2.

- **Symmetric Mean Absolute Coverage Error (sMACE)**: An indicator of the overall magnitude of calibration error.

$$\text{sMACE} = \frac{1}{n} \sum |\hat{c} - c| \,. \tag{5.1}$$

- **Asymmetric Mean Absolute Coverage Error (aMACE)**: An indicator of the overall magnitude of calibration error, where undercoverage (failure to meet nominal coverage) is penalised more than overcoverage (conservativeness).

$$\text{aMACE} = \frac{1}{n}\sum \left[\lambda_u \max(c - \hat{c}, 0) + \lambda_o \max(\hat{c} - c, 0)\right], \quad \text{where} \quad \lambda_u = 2, \lambda_o = 1. \tag{5.2}$$

- **Overcoverage Mean Absolute Coverage Error (oMACE)**: An indicator of how conservative a conformal method is.

$$\text{oMACE} = \frac{1}{n}\sum \max(\hat{c} - c, 0)\,. \tag{5.3}$$

- **Undercoverage Mean Absolute Coverage Error (uMACE)**: An indicator of how often a conformal method fails to meet nominal coverage.

$$\text{uMACE} = \frac{1}{n}\sum \max(c - \hat{c}, 0)\,. \tag{5.4}$$

- **Mean Average Prediction Set Size (APSS)**: An indicator of the overall efficiency of the conformal method.

$$\text{Mean APSS} = \frac{1}{n}\sum \text{APSS}. \tag{5.5}$$

The following datasets are used: Multiple-choice question-answering (MCQA) MMLU [167] and MedMCQA [168], open question-answering (OpenQA) TriviaQA [169], CoQA [170], and Natural Questions [171], mathematical reasoning GSM8K [172] and MATH [173], visual question-answering (VQA) MMMU [174], AI2D [175], and ScienceQA [176], and long-form generation MIMIC-CXR [177], FActScore [178], MedLFQA [79,179], and Wikipedia Biographies [65].

6. Trends in the Research

After systematically synthesising key characteristics of all conformal prediction studies featured in this review, we conduct frequency analysis and observe patterns and trends in this growing field, with respect to large language models in Section 6(a) and tasks & datasets in Section 6(b).

(a) Large Language Models

Across the 106 papers surveyed in this review, a wide variety of LLMs were utilised. In total, 424 LLM instances were evaluated, comprising 224 unique models across 42 model families. There were 31 LLM providers in total, though the majority of model families were distributed across the top seven providers, illustrated in Figure 3.

The top providers were dominated by American and Chinese technology or AI companies, which developed 50% and 37% of all LLMs evaluated in studies surveyed in this review, respectively. In comparison, the remaining 6% and 7% are developed in Europe and elsewhere, respectively. The leading provider in Figure 3 is Meta, popular for its open-source LLaMA family of models, followed by the Chinese open-source leader Alibaba, responsible for the versatile Qwen series, and OpenAI, the company behind ChatGPT. The distribution of models tracks the frequency of the top providers; the LLaMA model family [180–182] was the most popular across the literature (32%), followed by Qwen [183–186] (17%) and GPT [187,188] (15%), then Mistral and Mixtral [189,190] (6%), DeepSeek [191,192] (5%), and Gemini [193–195] (4.5%).

We also quantify the divide between open-source (i.e. white-box LLMs; public model weights, logit access) vs. closed-source (black-box LLMs; no public weights or logits, API-only access) across the papers we surveyed (see the pie chart in Figure 3). We found that the majority (81%) of studied models were open-source, possibly due to permissive licensing, widespread availability, and lower-cost inference. Closed-source models were the minority (19%), and mostly comprised OpenAI's GPT-4.x and GPT-4o and Google's Gemini. Unfortunately, the lack of frontier model

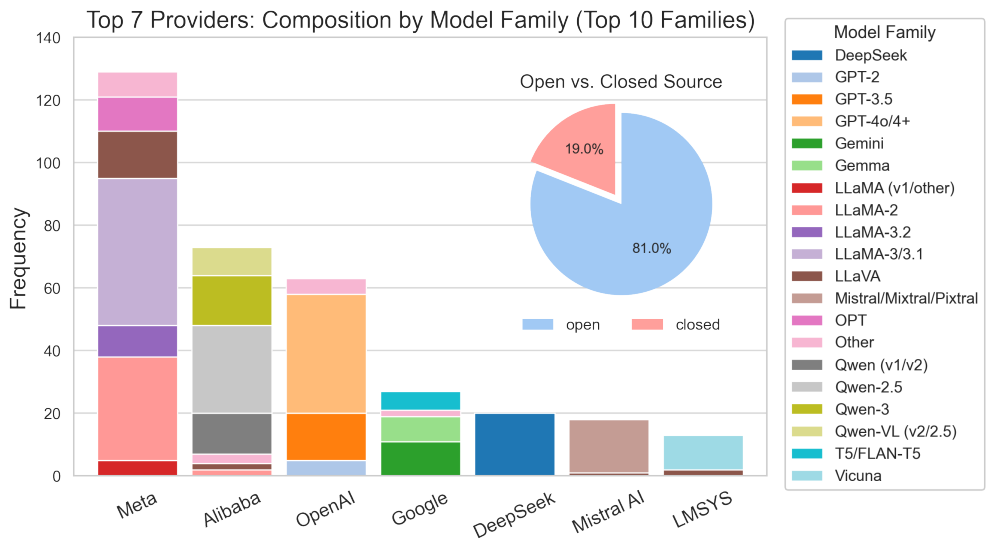

Figure 3: We find that the majority of top providers are American and Chinese technology or AI companies, with Meta leading due to widespread use of its high-performance, low-cost open-source LLaMA family of models. In addition, we find that only 19% of the LLMs studied are closed-source, indicating a research gap with proprietary and frontier LLMs validation.

representation is a gap in the literature, as the research into uncertainty quantification with rigorous statistical guarantees of validity for the most popular LLMs is limited.

Research has shown that even advanced LLMs still suffer from hallucinations, bias, and overconfidence [4–6]. Whilst frontier models are often proprietary, open science and the dissemination of ideas encourage rapid innovation, with companies like Meta, Alibaba, and DeepSeek paving the way for industry-academic collaboration. Whilst improving the reliability of white-box open-source models is a worthy endeavour, frontier LLMs such as ChatGPT, Claude, and Gemini are used by millions of non-technical consumers every day [196–198], and are now the powerhouses behind many businesses [199]. Therefore, we can say that improving the reliability of these frontier models will have the greatest impact on human society.

Recommendation. With this in mind, we encourage researchers in this field to continue to pursue uncertainty quantification techniques for black-box models with conformal prediction. Whilst several logit-free methods have been proposed [69,73,75,98,106], there are still many challenges to overcome, discussed in more detail in Section 13. Many logit-free methods are validated on GPT-3.5, which is no longer representative of the state-of-the-art. We therefore recommend that researchers continue to adapt conformal prediction to black-box models, and validate on a more diverse selection of closed-source models, more frequently. We recommend a split of 40%/60% closed-source and open-source. For example, no Anthropic LLM is included in the 224 unique models evaluated in this field so far, despite widespread consumer popularity.

(b) Tasks and Datasets

Beyond LLM model families, we also analysed the distribution of tasks and the evaluation datasets used in all 106 of the papers we surveyed in this review. There were 53 distinct tasks identified overall, with 197 unique datasets that spanned these tasks. Figure 4 illustrates the top 30 datasets and their task-wise compositions with the top 16 tasks overall.

The two most popular datasets in the conformal prediction for LLMs literature are TriviaQA [169] for open-domain QA (6.8%) and MMLU [167] for multiple-choice QA (5.9%), as shown

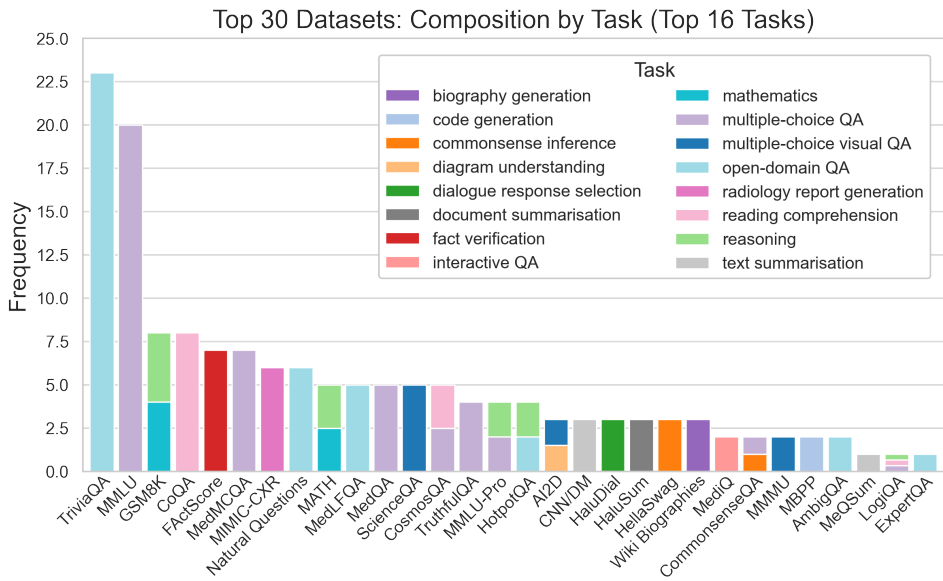

Figure 4: The task composition of the top 30 datasets was relatively diverse, with the top 16 tasks consisting of most open-domain QA (15%), multiple-choice QA and reasoning (both 11%), followed by mathematics and reading comprehension (both 6%), then fact verification (3.8%) and common-sense inference (2.6%). The remaining tasks include radiology report generation, text summarisation, and multiple-choice visual QA, among others, each task with a share of ≈2.3%.

in Figure 4. The remaining 195 datasets are spread widely, with many datasets used only once or twice in total. Following on from TriviaQA and MMLU, the remaining eight datasets in the top ten are GSM8K and CoQA [170,172] (2.4%), FActScore and MedMCQA [168,178] (2.1%), MIMIC-CXR and Natural Questions [171,177] (1.8%), and MATH and MedLFQA [173,179] (1.5%). These datasets are heterogeneous in their task composition; GSM8K and MATH are for mathematical reasoning, CoQA is a QA dataset for reading comprehension, FActScore is for fact verification, and Natural Questions and MedMCQA are open-domain QA and multiple-choice QA, respectively. MedLFQA is for open-domain long-form QA, often used in conjunction with FActScore for evaluating factuality. MIMIC-CXR is for radiology report generation (RRG).

Prior to conducting our controlled performance analysis, we utilised frequency analysis to make our dataset selections, ensuring that we achieved a diverse set of representative tasks. We took the top ten datasets and then aggregated them with the top three multi-modal datasets, ScienceQA [176] (1.5%), AI2D [175] (0.9%), and MMMU [174] (0.6%), for multiple-choice visual QA and diagram understanding tasks to evaluate LVLMs. Lastly, we added the Wikipedia Biographies [65] (0.9%) dataset to further diversify our task selection, and because it was a popular benchmark for factuality evaluation, in particular conditional coverage, a test setting we wanted. Overall, our selection of 14 datasets for the controlled performance analysis covered a diverse range of tasks, including OpenQA, MCQA, mathematical reasoning, RRG, biography generation, reading comprehension, fact verification, MCVQA, and diagram understanding.

Overall, observing the full dataset distribution and the conformal methods are evaluated on it, we find some interesting trends. Although the literature is characterised by substantial heterogeneity and an abundance of datasets, conformal studies tend to fall into one of two dichotomous categories; some rely on relatively obscure datasets that have not been used in prior work in this field, or introduce new datasets tailored specifically to the task or application they are researching (such is common in robotics studies [52,83,85,96,120,159]). Others fall into the second category, evaluating their methods on a small set of widely used benchmarks in their

niche. For example, most split-conformal calibration papers are evaluated on TriviaQA or MMLU, depending on whether they focus on open- or closed-ended calibration (thus, we get the dataset inflation observed for both in Figure 4).

Recommendation. Considering this, we suggest that researchers in this field take a blended approach; consistent evaluation on well-designed publicly released benchmarks such as TriviaQA should be encouraged, but considerations of task diversity should be made. We also recommend evaluating on a minimum of four datasets to show generalisability. Consider the following,

- **EX.1.** From evaluating on MMLU, MedMCQA, and MedQA [200], which are all multiple-choice, domain-specific QA datasets, to MMLU/MedMCQA, CosmosQA [201], CommonsenseQA [202], and LogiQA [203] for a more diverse task selection, as evaluation now covers domain-specific MCQA, general knowledge MCQA, reading comprehension, commonsense inference, and disjunctive reasoning. If the focus must be on the medical domain, extending the evaluation to include a dataset such as AfriMedQA [204], a pan-African dataset with localised clinical examples absent in other medical benchmarks, could provide a more rigorous assessment of robustness and bias.
- **EX.2.** From evaluating on TriviaQA, Natural Questions, and PopQA [205], which are all open-domain general knowledge QA datasets of varying difficulty, to TriviaQA, CoQA, HotpotQA [206], and AmbigQA [207] for a diversified task selection including open-domain general knowledge QA, multi-step reasoning, reading comprehension, and question disambiguation, as well as QA of varying difficulty.

Although, for more specialised research areas, such as conformal factuality, the choice of dataset is narrower, there are still applicable options. For example, a factuality validation pipeline with MedLFQA, Wikipedia Biographies, and FactScore, can be further augmented with ExpertQA, a long-form open-domain QA dataset with expert-curated questions across 32 specialised fields [208]. For studies evaluating a conformal abstention mechanism, datasets such as AmbigQA or TruthfulQA [209], which include ambiguous or adversarial questions, can be useful to test abstention vs. hallucination rates; models which cannot disambiguate questions should choose to abstain, or models should abstain rather than mimic human falsehoods (in the case of TruthfulQA). For conformal RAG, PopQA for testing parametric knowledge should be utilised.

7. Split Conformal-based Methods

This section explores methods based on the split conformal prediction framework[3] outlined in Section 4. It covers methods in the open-ended and close-ended calibration branch nodes in Figure 2. The overall section is organised as follows. Section 7(a) surveys logit-based methods for LLMs and LVLMs, Section 7(b) reviews logit-free methods for LLMs and LVLMs, and Section 7(c) presents token-level uncertainty methods. The performance of all methods is reported in Table 3.

(a) Logit-based Methods

This section is organised as follows. Section i surveys methods applied to close-ended tasks, Section ii discusses a performance comparison between LLMs and LVLMs, Section iii presents optimisation frameworks for close-ended tasks, Section iv reviews methods for open-ended tasks, and Section v discusses distribution-shift-aware methods.

(i) Close-ended Split-conformal Calibration

Close-ended calibration applies to tasks where a language model's output is restricted to a predefined set of possible responses, as shown in Figure 5. Such tasks include multiple-choice

[3]With one exception; conformal language modelling is built upon the LTT framework, but as we do not have a dedicated LTT section and it is the first conformal method proposed for open-ended tasks, we place it in Section iv instead.

question-answering (MCQA), text classification, and sentiment analysis. The following studies explore the application of conformal prediction to close-ended tasks.

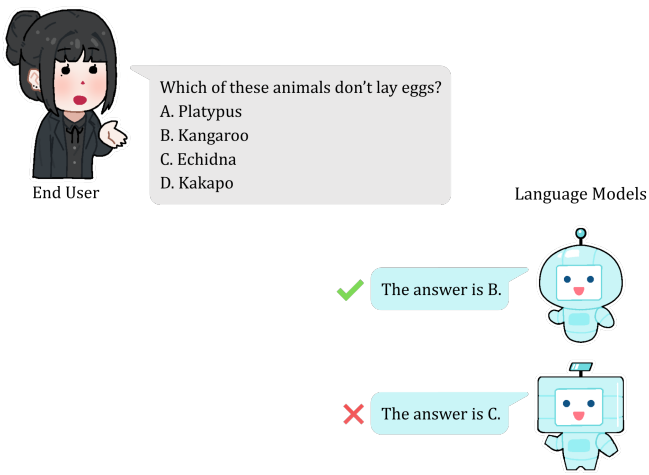

Figure 5: **Close-ended Calibration** considers tasks where the language model's output is constrained to a predefined set of responses. A typical example is multiple-choice question-answering, where the model selects from n answer options; four options in this case. The green check mark indicates that the answer is correct, and the red cross indicates that it is incorrect.

Conformal prediction router (CP-Router) is proposed to address an efficiency problem; large reasoning models (LRMs) often outperform LLMs on difficult questions, but they can also "overthink" easy prompts, generating many more tokens than necessary and sometimes harming accuracy. CP-Router is a training-free, model-agnostic routing mechanism that decides when a prompt should be answered by a lightweight LLM and when it should be escalated to a more expensive LRM [99]. Methodologically, the router is built around conformal prediction in the MCQA setting. For each prompt, the LLM's option probabilities are converted into nonconformity scores $1 - f(y)$, calibrated on a held-out set, and used to form a prediction set whose size serves as an uncertainty signal; small sets are routed to the LLM, while larger sets are sent to the LRM. To avoid hand-tuning the error level α, full and binary entropy (FBE) is introduced, which selects α by maximising both the diversity of prediction set sizes and the balance between singleton and non-singleton sets. Experiments across several MCQA benchmarks and an adapted open-ended question-answering (OpenQA) setting show reduced token usage with accuracy comparable to, and sometimes better than, always using the LRM, supporting the utility of CP-Router.

LLMs for interactive question-answering, where the model must decide which question to ask next in order to make an accurate prediction with as few turns as possible, remained unaddressed in the literature. To address this, conformal information pursuit (C-IP) is proposed, which replaces entropy-based uncertainty in standard information pursuit (IP) with the average size of conformal prediction sets [107]. The central insight is that conditional entropy can be upper bounded by a function of expected prediction set size, so query informativeness can be approximated by choosing the next query that minimises this conformal-set-based bound. C-IP constructs prediction sets using the split-conformal framework over histories of query-answer chains, with two strategies for sampling histories:tight uniform sampling over a closed query set and LLM-simulated histories for open-ended querying. Experiments on 20 Questions show that C-IP typically yields more informative uncertainty estimates, shorter query chains, and better predictive accuracy than entropy-based IP and direct prompting; on the MediQ interactive medical QA dataset, it achieves competitive performance with single-turn prediction while providing interpretable QA trajectories. The main limitation is that C-IP uses marginalised

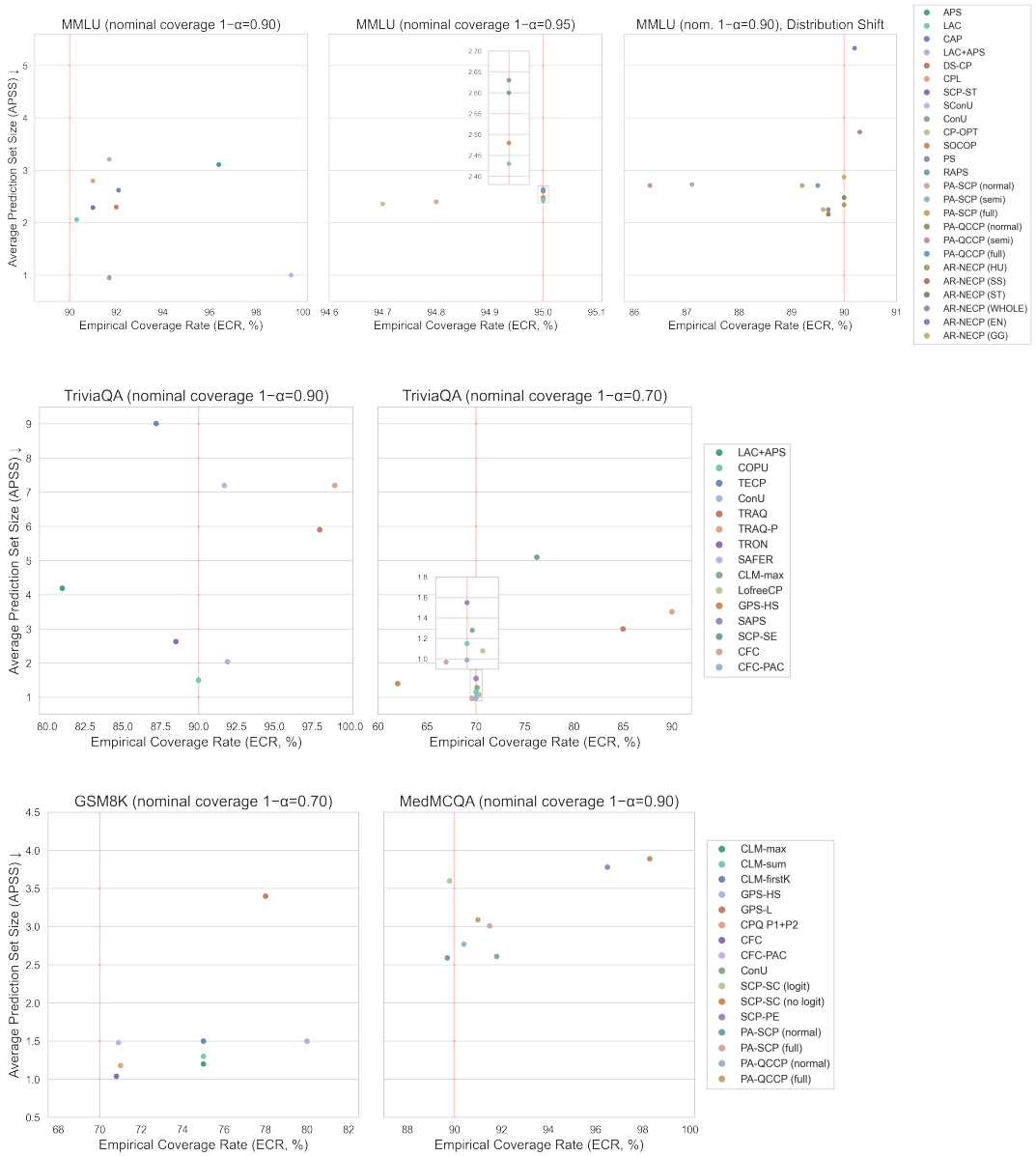

Figure 6: Scatterplots for split-conformal-based methods on MMLU, TriviaQA, GSM8K, and MedMCQA. Marginal coverage at various targets is explored, including scenarios under distribution shift. Metrics are defined in Section 5. The x-axis is empirical coverage. The red line indicates the target coverage; to the left of this line is undercoverage and to the right of this line is overcoverage. The y-axis is average prediction set size; lower values are better.

coverage over query-chain lengths rather than guarantees for the exact realised history. Coverage can also underperform when calibration data is small or the query space is too broad.

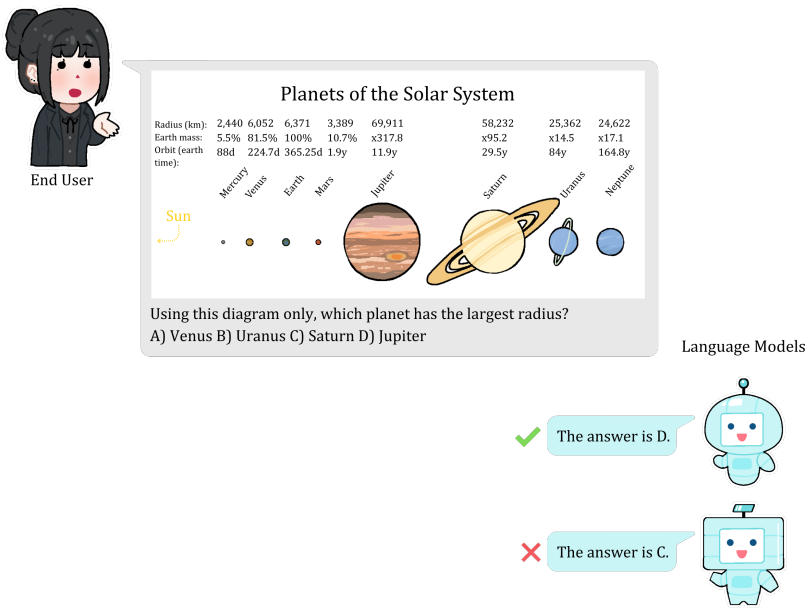

Figure 7: **Close-ended Calibration** also applies to visual tasks, such as multiple-choice visual question-answering, where the output space is restricted to a fixed set of answer options. In this setting, the LVLM must use visual information from the image to infer the correct answer. The green check mark indicates that the answer is correct, and the red cross indicates it is incorrect.

(ii) Performance Comparison of LVLMs and LLMs

As with text-based MCQA, MCVQA is a close-ended task where the language model's output is restricted to a predefined set of possible responses, as shown in Figure 7. In the experiments outlined in Section 5, several conformal methods for LVLMs are evaluated in Figure 8.

In the review objectives outlined in Section 1(a), we proposed **Hypothesis B**. To recap, we hypothesised that conformal methods applied to LVLMs may face a more challenging calibration problem than those applied to the text-only setting of LLMs, due to additional sources of uncertainty arising from visual perception, cross-modal fusion, and ambiguity in grounding text to image regions. We posit that we would **see a higher coverage error, particularly overcoverage error, and larger and less informative prediction sets in vision-language tasks**.

In Section 5, we outlined a controlled performance analysis, from which we conducted additional subgroup studies using a diverse subset of those methods. One such subset was divided based on whether a conformal method was applied to LLMs or LVLMs. We then computed four variants of mean absolute coverage error (MACE) and mean average prediction set size (APSS) as defined in Section 5, and report our results in Figure 11, under 'LLM vs LVLM' and the bars with purple hue. We find that overall, conformal methods applied to LVLMs exhibited a moderately lower symmetric MACE (sMACE) and asymmetric MACE (aMACE), and significantly lower undercoverage error, whereas conformal methods applied to LLMs had a much higher undercoverage error. We also observe a slight increase in APSS for LVLM methods, and find that they tend to be more conservative on average, exhibiting a marginally higher overcoverage error than LLM methods. The observations we made partially supported our initial hypothesis. Whilst we observed a lower coverage error for LVLM methods rather than what we hypothesised, we did correctly anticipate that LVLM methods would have both a higher overcoverage error and APSS on average than LLM methods. In addition, the near-zero undercoverage error is likely to be what lowered the sMACE and aMACE; especially aMACE, as this penalises failure to meet the nominal coverage more heavily than conservativeness.

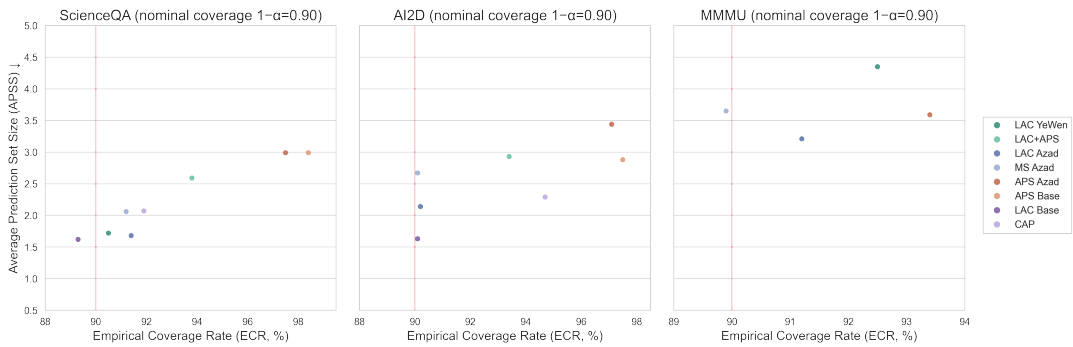

Figure 8: Scatterplots for split-conformal-based methods on ScienceQA, AI2D, and MMMU with LVLMs at marginal coverage. Metrics are defined in Section 5. The x-axis is empirical coverage. The red line indicates the target coverage; to the left of this line is undercoverage and to the right of this line is overcoverage. The y-axis is average prediction set size; lower values are better.

In the literature, LVLMs have been widely reported to suffer from visual hallucinations, especially pronounced in open-ended generation and stemming from failures tied to visual grounding and perception, modality interaction, and ambiguous or anomalous inputs [94, 124,126,210–212]. In many cases, accuracy in LVLMs is not tightly aligned with uncertainty, such that stronger performance does not always translate to sharper and less conservative conformal behaviour [124,212]. Open-ended multimodal tasks are shown to introduce semantic redundancy and instability, which can enlarge prediction sets and reduce their informativeness [69]. In addition, in some domains such as mathematics and reasoning-heavy multimodal tasks, performance worsens, which suggests that multimodal uncertainty is heterogeneous and domain-sensitive, which could manifest as conservative conformal behaviour [126].

Therefore, the literature is consistent with our claims that LVLMs may have additional sources of uncertainty arising from the visual encoder and multimodal fusion components. However, the mechanisms underlying our results are not fully clear. On the other hand, these additional uncertainty signals broaden conformity scores or make them noisier, encouraging conformal methods to opt for larger and safer prediction sets; this may explain the higher overcoverage error and APSS, and lower undercoverage error we observed. However, the literature also finds that open-ended generation is a major uncertainty contributor [69,211], whereas all of the LVLMs in our controlled performance analysis, with the exception of one, were evaluated on close-ended tasks, i.e. multiple-choice visual question-answering (MCVQA).

Therefore, while we do not claim a definitive causal explanation, the empirical evidence indicates that, in our experiments, **conformal methods for LVLMs exhibit a stronger coverage-informativeness trade-off than those for LLMs**.

(iii) Close-ended Split-conformal Calibration with Set Optimisation

Two common primary objectives for split-conformal based methods for discrete LLM tasks, such as MCQA, are guaranteeing marginal coverage and decreasing the length of prediction sets. However, often the former takes priority with the latter being considered a secondary criterion, which limits the informativeness of the prediction sets in downstream decision-making due to unnecessarily large sizes. In addition, marginal coverage can often be too weak when one needs validity across subpopulations or under structured covariate shifts. Therefore, conformal prediction with length-optimisation (CPL) is proposed as a principled framework for minimising expected prediction set length subject to conditional validity constraints [82].

It is well known that Conformal Prediction has two major challenges in practice: conditional validity and length inefficiency. Conditional validity means for every $x \in \mathcal{X}$, $\mathbb{P}(y_{n+1} \in$

$C(x_{n+1})|x_{n+1} = x) = 1 - \alpha$ for a given miscoverage rate α. On the other hand, length efficiency is about producing prediction sets whose length is as small as possible while maintaining the required conditional validity. The length objective is the expected size of the prediction set $C(x)$, which is cardinality in the classification setting. The optimisation problem is therefore for a given miscoverage rate α, $\min_C \ \mathbb{E}[\text{len}(C(x))] \quad \text{s.t.} \quad \mathbb{E}\big[f(x)\big(\mathbf{1}(y \in C(x)) - (1-\alpha)\big)\big] = 0, \ \forall f \in \mathcal{F}$, where $\mathbf{1}(.)$ is an indicator function, and $\mathcal{F}$ is a class of functions from $\mathcal{X}$ to $\mathbb{R}$ to model covariate shift. The first step is to define the Lagrangian-like objective, $g_\alpha(f, C) = \mathbb{E}\big[f(x)\big(\mathbf{1}(y \in C(x)) - (1-\alpha)\big)\big] - \mathbb{E}[\text{len}(C(x))]$ yields $\min_{f \in \mathcal{F}} \max_C \ g_\alpha(f, C)$. A strong duality result is proven; under continuity assumptions, the above minimax problem is equivalent to the "minimum-length subject to conditional validity" problem. Moreover, if (f^*, C^*) is optimal, then $C^*(x) = \{y \in \mathcal{Y} : f^*(x)p(y \mid x) \geq 1\}$. The optimally conditionally valid prediction set is a covariate-adaptive density level set. As such, length-optimal prediction sets should look like thresholded level sets of $p(y \mid x)$, where the threshold is adapted through $f^*(x)$ to satisfy the desired validity constraints. In the marginal-coverage special case, this reduces to a fixed global density level set.

For a given dual variable f, the best set is exactly the corresponding density level set. The outer minimisation over f then chooses the one that enforces the desired validity constraints over the class $\mathcal{F}$. This is interpreted as follows: the outer minimisation navigates the space of conditionally valid sets, while the inner maximisation picks the shortest one within that direction. Because the optimal sets depend on the unknown density $p(y \mid x)$, a more practical relaxed minimax formulation based on a given conformity score $s(x, y)$ is given. Instead of optimising over arbitrary sets, it restricts to structured sets of the form $C_h^s(x) = \{y \in \mathcal{Y} : s(x, y) \leq h(x)\}$ and optimises over threshold functions $h \in \mathcal{H}$, where s is any chosen conformity score and $\mathcal{H}$ is a class of real-valued functions on the set $\mathcal{X}$. The score is not optimised, only the adaptive threshold function $h(x)$ on top of a fixed score. This is the distinction from works such as [213]. CPL then solves the smoothed finite-sample relaxed minimax problem by alternating ascent on h and descent on f. The main finite-sample result shows that for any $f_\beta \in \mathcal{F}$ when f_β is represented by a $\beta \in \mathbb{R}^d$ and $f(x) = \langle \beta, \Phi(x) \rangle$, the conditional-coverage violation of the learned set is bounded.

In the experiments outlined in Section 5, CPL is evaluated in Figure 6a. On MMLU at a target marginal coverage of 90%, CPL meets the target and achieves an APSS of 2.80, outperforming APS-based methods. The paper does not report results for CPL on conditional coverage targets. On MMLU at a target marginal coverage of 95%, CPL is just shy of the target by 0.2%, and is ranked 2$^{\text{nd}}$ in terms of the lowest APSS out of three nonconformity score functions and two other optimisation methods evaluated. CPL only guarantees conditional validity relative to a user-chosen finite-dimensional class $\mathcal{F}$ of covariate shifts. Note that the guarantees are only as strong as the chosen feature basis $\Phi(x)$. If subpopulation structure is not represented in $\mathcal{F}$, CPL may still miscalibrate on those hidden subgroups. Additionally, in finite samples, the guarantee is only approximate conditional validity at stationary points of the smoothed minimax objective, not exact optimality of the algorithm. The coverage error is shown to decrease at a rate $O(n^{-1/2})$, but there is no proof that the learned stationary point achieves finite-sample length optimality.

Many conformal set optimisation methods for close-ended LLM tasks such as MCQA aim to minimise APSS, but in many practical settings the most valuable outcome is a singleton set, because non-singleton outputs often require additional human intervention or are associated with downstream ambiguity; for example, in MCQA, a single, correct answer is objectively more useful to the end user than a set of possible, but not explicitly correct answers. Motivated by the need to optimise conformal methods not merely for shorter answer sets on average but for a higher frequency of unambiguous answers, singleton-optimised conformal prediction (SOCOP) is proposed [146]. SOCOP introduces a new nonconformity score derived from a constrained optimisation problem that balances the probability of producing non-singleton sets against expected set size under a coverage constraint. By studying the Lagrangian form of this problem, it is shown that optimal prediction sets take the form of top-k label sets and are nested as the penalty parameter varies. This yields a conformal score that can be computed efficiently through a geometric reformulation; the key step reduces to finding the lower convex hull of

K two-dimensional points, enabling $O(K)$ computation for K-class problems. The resulting split-conformal procedure preserves standard marginal coverage guarantees. In the experiments outlined in Section 5, SOCOP is evaluated in Figure 6a. On MMLU at a target coverage of 95%, SOCOP is ranked 4$^{\text{th}}$ in terms of the lowest APSS out of all nonconformity functions and two optimisation methods whilst maintaining tight coverage. However, SOCOP is not optimised for APSS, but for the singleton rate (i.e. reducing the NSR). Whilst not enough methods reported NSR for it to be considered as a metric in the experiments outlined in Section 5, Table 3 report the NSR for several SCP methods, in which SOCOP is ranked 4$^{\text{th}}$ irrespective of user-specified error level α, and ranked 1$^{\text{st}}$ when fixing $\alpha = 0.05$ (target coverage 95%). A limitation of SOCOP is that it targets marginal, not conditional, coverage, and it requires tuning a regularisation parameter λ, which shapes the trade-off between singleton frequency and average set size.

(iv) Open-ended Split-conformal Calibration

Whilst these studies demonstrate the utility of split-conformal-based uncertainty quantification for LLMs, they have narrow experimental scopes, focusing on close-ended tasks for LLMs such as binary QA, MCQA, and text classification, where the output is constrained within a predefined set of possible answers. A more challenging task is that of open-ended generation, where the output is not predefined and there is a wide range of semantically-equivalent acceptable responses, and there may be no unique correct response. Some examples of open-ended tasks for LLMs include text summarisation and open-domain question answering (OpenQA), as shown in Figure 9. The following studies explore the application of conformal prediction to open-ended tasks.

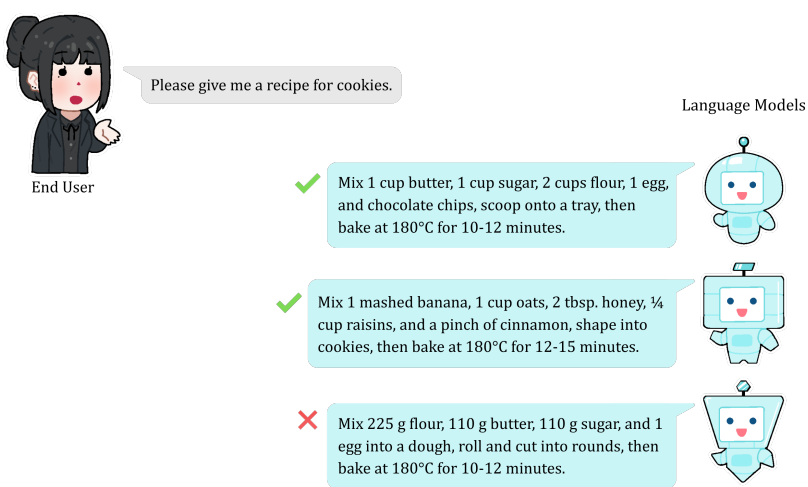

Figure 9: **Open-ended Calibration** considers tasks where the language model's output space is unbounded. Multiple responses may be semantically equivalent and valid, and there may be no unique correct answer. A typical example is open-domain question-answering, where the model responds with free-form text rather than selecting a predefined option. The green check mark indicates that the answer is correct, and the red cross indicates that it is incorrect.

The first method to extend conformal prediction from classification to open free-text generation is conformal language modelling (CLM) [53]. The authors note the differences between classical conformal prediction building prediction sets by testing candidate labels in a finite or otherwise tractable output space, and how this cannot be done for open generation where the output space is effectively unbounded and combinatorial. Rather than filtering over an intractable output space, CLM calibrates a sampling algorithm called conformal sampling with rejection. Given an

LLM, a set-confidence function F (indicating that the retained set contains a sufficiently high probability of including at least one acceptable answer.), a similarity function S (e.g. ROUGE-L), and a quality function Q (e.g. the LLM's likelihood function), the algorithm repeatedly samples outputs y_k from the LLM, rejects low-quality samples $Q(x, y_k) < \lambda_2$, rejects near-duplicates $\max_{y_j \in C_\lambda(x)} S(y_k, y_j) > \lambda_1$, and stops once $F(C_\lambda(x)) \geq \lambda_3$ where $\lambda = (\lambda_1, \lambda_2, \lambda_3)$ are our choice of hyper-parameters. There is a calibration set $D_{\text{cal}} = \{(x_i, A_{x_i})\}_{i=1}^n$ where x_i is a prompt and $A_{x_i}(y)$ is a binary function that measures whether or not a generation $y \in \mathcal{Y}$ for prompt x_i is "good enough" (i.e. $A_{x_i}(y) = 1$ means y is acceptable). Using the calibration set, the objective is to determine the choice of hyperparameters $\lambda = (\lambda_1, \lambda_2, \lambda_3)$ and to build a sampled prediction set $C_\lambda(x_{\text{test}}) \subseteq 2^{\mathcal{Y}}$ for a new prompt x_{test} satisfying

$$\mathbb{P}\Big(\mathbb{P}\big(\exists\, y \in C_\lambda(x_{\text{test}}) : A_{x_{\text{test}}}(y) = 1 \mid D_{\text{cal}}\big) \geq 1 - \varepsilon\Big) \geq 1 - \delta, \tag{7.1}$$

where ε is the tolerated test risk and δ controls the dependence on the specific calibration draw. The calibration uses LTT [36], and for each configuration λ, the prediction set $C_\lambda(x_i)$ is constructed by sampling until a calibrated level of confidence that a good answer is included is reached. The corresponding loss is $\mathcal{L}_{x_i}(\lambda) = \mathbf{1}(\nexists y \in C_\lambda(x_i) : A_{x_i}(y) = 1)$, with the empirical risk $\hat{R}_n^{\text{b}}(\lambda) = \frac{1}{n}\sum_{i=1}^n \mathcal{L}_{x_i}(\lambda)$, and a valid Binomial-tail p-value for testing the null hypothesis $\mathcal{H}_\lambda$: $\mathbb{E}[\mathcal{L}_{x_{\text{test}}}(\lambda)] > \varepsilon$ is $p_\lambda^{\text{BT}} = \mathbb{P}\big(\text{Binom}(n, \varepsilon) \leq n\,\hat{R}_n^{\text{b}}(\lambda)\big)$, where $\text{Binom}(n, \varepsilon)$ denotes a binomial random variable with sample size n and success probability ε. After applying a multiple-testing correction, the selected valid configuration is the one that minimises a weighted trade-off between the final set size and the additional sampling cost required to achieve the target confidence level.

They also introduce component-level selection, where a component $e \in \mathcal{Y}$ is a logically defined subpart of a larger response y. $E(y)$ is a deterministic function that takes a response y and breaks it down into components. For every input x, there is some component-based admission function $A^c(e)$, which can be used to judge individual components e for correctness. $F^c(e)$ is a function that gives a confidence score for the event $A^c(e) = 1$ for component e. The subset of components for a threshold γ is defined as $C_\gamma^{\text{inner}}(x) = \{e \in \cup_{y \in C_\lambda(x)} E(y) : F^c(e) \geq \gamma\}$. Using the calibration set with a separate LTT calibration, $\hat{\gamma}$ can be estimated to guarantee that for test pair $(x_{\text{test}}, A^c_{x_{\text{test}}})$ the selected components $C_{\hat{\gamma}}^{\text{inner}}(x_{\text{test}})$ are correct with probability at least $1 - \alpha$. This means that the CLM method not only provides a set-level guarantee that at least one sampled response is acceptable, but also a way to identify reliable sub-components such as sentences inside longer generations. This is motivated by the idea that even when a fully generated answer is imperfect, parts of it may still be independently correct.

CLM has some limitations. Firstly, its dependence on an admission function A that must be a meaningful proxy for true answer quality. In the experiments, automatic metrics define admissibility, and so the guarantee is therefore only as meaningful as the admission function. Secondly, the approach may become impractical if the sampled sets are too large or too expensive to obtain. Because of the dependence on repeated LLM sampling, CLM requires that acceptable responses occur with sufficient probability under the base model. This means that good outputs must exist and must be sampled often enough in a tractable number of calls, otherwise a valid configuration will not be found and CLM will abstain.

In the experiments outlined in Section 5, CLM with `max`, `sum`, and `first_k` scoring (denoted CLM-max, CLM-sum, and CLM-firstk, respectively) are evaluated in Figure 6c, and in Figure 6b, only CLM-max is evaluated. On GSM8K at a target coverage of 70%, CLM with all scores had an overcoverage of $\approx 5\%$, indicating the method is somewhat conservative. In terms of APSS, CLM-max outperformed the other two scores, with CLM-firstK yielding the highest APSS out of the three. On TriviaQA at a target coverage of 70%, CLM-max ranked 4$^{\text{th}}$ in terms of lowest APSS out of the seven other methods evaluated, and maintained tight coverage.

A primary critique of CLM is that it builds prediction sets by sequentially examining generated samples and calibrating three thresholds $(\lambda_1, \lambda_2, \lambda_3)$ through the LTT framework. This filtering scheme can make CLM both computationally heavier and more fragile, as it wrongly rejects correct samples which inflates the effective abstention rate. Thus, for a tight α, CLM may fail

to find a valid configuration at all and must abstain. In addition, the abstention is binary at the calibration level; if it finds a valid configuration, it will never abstain, and if not, it will always abstain. Another conformal method, generative prediction sets (GPS), is proposed to address this limitation [101]. GPS formalises an admissibility function that deems a generated output acceptable if it satisfies an application-specific criterion, such as passing unit tests in code generation. GPS reduces prediction set construction to conformal regression on the minimum number of samples required to obtain an admissible output. For each calibration input, they estimate this stopping time up to a fixed sampling budget, treat failures as a sentinel $M+1$, and conformalise a predictor of the success probability using one-sided CQR (see Section 4).

Prior logit-free sampling-based conformal methods for open-ended generation [73,75] (see Section 7(b)) rely on the restrictive assumption that a correct answer already appears within the sampled candidate set, which may not hold in realistic scenarios. Conformal prediction for uncertainty quantification, or COPU, is a reprompting-based split-conformal framework [145]. For each input, the LLM first samples K unique candidate responses; if the ground truth answer is absent, it is explicitly injected into the candidate set. The LLM is then reprompted with the original input concatenated with each candidate answer, and token-level conditional probabilities are multiplied to obtain a sequence score. After softmax normalisation, the least ambiguous classifier-style nonconformity score is defined and calibrated on a held-out set to produce prediction sets with target coverage. In the experiments outlined in Section 5, COPU is evaluated in Figure 6b. On TriviaQA at a target coverage of 90%, COPU achieves the lowest APSS out of the six other methods and their variants, whilst maintaining tight coverage. On TriviaQA at a target coverage of 70%, COPU achieves the 4^{th} lowest APSS out of the eight other methods and their variants, whilst maintaining tight coverage. The limitations of COPU are that it requires logit access whereas the methods it critiqued did not, it approximates the full sequence space through finite sampling, and its performance depends on the chosen nonconformity function.

(v) Distribution-shift-aware Methods

Standard conformal prediction has a well-known weakness; its finite-sample validity relies on exchangeability between calibration and test data. For deployed LLMs, this assumption is often violated because prompts at deployment time originate from domains that differ from the calibration set (known as distribution shift, see Figure 10). Conformal prediction methods for LLMs have been shown to be under-cover from domain shift, producing prediction sets that are too small and therefore falsely reassuring to the end user. In addition, standard covariate-shift weighted conformal prediction is not directly suitable for LLM prompts, because prompts are high-dimensional and unstructured inputs, so direct density-ratio estimation (DRE) in the raw prompt space is computationally infeasible. The proposed method, domain-shift-aware conformal prediction (DS-CP), aims to represent prompt similarity in a semantically meaningful lower-dimensional space [132].

There is a dataset of prompt–ground truth pairs $\{(x_i, y_i)\}_{i=1}^{n}$ from the old domain, which will be used as calibration data. For the new domain, only prompts are available and their truth labels are not observed. DS-CP first maps prompts into a lower-dimensional semantic space using a pre-trained embedding model $h : \mathcal{X} \to \mathcal{Z}$. Sentence embeddings from a MiniLM SentenceTransformer are used so that prompts with similar meaning are mapped nearby; this is intended to capture cross-domain similarity more effectively than raw prompt tokens. The density ratio is $r_e(z) = \frac{dP'_Z}{dP_Z}(z)$, where dP_Z and dP'_Z denote the old-domain and new-domain marginal distributions in embedding space. If this ratio were known, the weighted split-conformal framework [42] could be applied. However, even in embedding space the dimensionality remains high enough that estimated density ratios can be extremely imbalanced, in particular, if $r_e(h(x_{n+1})) \gg r_e(h(x_i)),\ \ i = 1, \ldots, n$ then the empirical score distribution places too much mass on the δ_∞ point used in weighted split-conformal, causing the prediction set to inflate towards all labels. To address this, DS-CP introduces a regularisation step; instead of using the test-point weight $r_e(h(x_{n+1}))$ directly, it replaces it with a regularised quantity $\lambda = \lambda(x_1, \ldots, x_n) \geq 0$, which

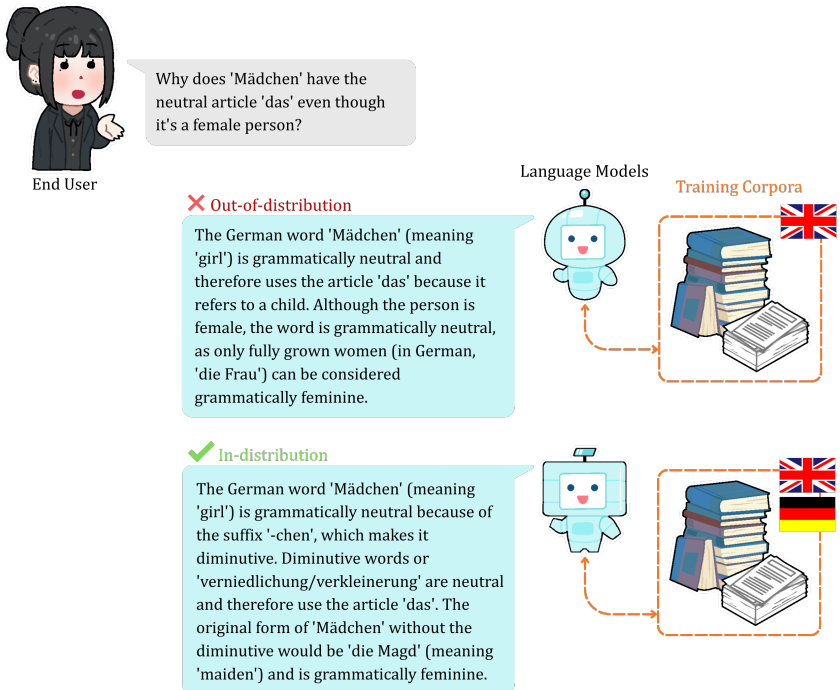

Figure 10: **Distribution Shift** in LLMs can arise when a prompt is out-of-distribution (OOD) relative to the model's training corpus. The model may lack reliable knowledge relevant to the query, yet still produce a fluent answer by extrapolating from weakly related patterns. This can lead to hallucination: the model generates content that is linguistically plausible but not grounded in true supporting information learned from pre-training. Here, a monolingual LLM fails to answer a multilingual question which is OOD, but the multilingual LLM does not. The green check mark indicates the correct answer, and the red cross indicates the incorrect hallucination.

depends only on the calibration data. Given the calibration scores $s_i = s(x_i, y_i)$ for $i = 1, ..., n$ and the estimated calibration weights $\widehat{w}_i = \widehat{r}_e(h(x_i))$, the resulting empirical distribution of scores is

$$\hat{\mu}_n = \sum_{i=1}^{n} \frac{\widehat{w}_i}{\sum_j \widehat{w}_j + \lambda} \delta_{s_i} + \frac{\lambda}{\sum_j \widehat{w}_j + \lambda} \delta_{\infty}, \tag{7.2}$$

where δ_s denotes the Dirac measure at $s \in \mathbb{R}$. The final prediction set for a test point x is $C(x) = \{y : s(x, y) \leq \text{Quantile}(\hat{\mu}_n; 1 - \alpha)\}$. To estimate the density ratio r_e, a domain classifier is trained in the embedding space; XGBoost [214] is used in this case. For the nonconformity score, the LAC [29] score is used, with ablations testing the APS [30] score as well. The main theory gives lower and upper coverage bounds involving the weights and total variation distances between calibration and test score distributions, which is more relaxed compared to the exact finite-sample, distribution-free guarantee of the split-conformal framework under exchangeability, but this is entirely reasonable under domain shift. When there is no shift and $\widehat{w}_i = \lambda = 1$, DS-CP reduces to standard split-conformal calibration. In the experiments outlined in Section 5, DS-CP is evaluated in Figure 6a. On MMLU at a target coverage of 90%, DS-CP achieves the joint 4$^{\text{th}}$ lowest APSS out of the eight other methods, and had a slight overcoverage of $\approx 2\%$. DS-CP has a few limitations; it depends on the quality of the embedding model and the density-ratio estimator. If the embedding fails to represent semantic domain similarity well, or if the domain classifier estimates poor ratios, the weights may be uninformative or unstable.

Whilst DS-CP is shown to be effective for close-ended MCQA, it is not designed for open-ended generation. Selective conformal uncertainty (SConU) [108], an extension of ConU [75] (see Section 7(b) for more details) is designed to address violations of exchangeability and uncertainty distribution shift between calibration and test data in for both close-ended and open-ended QA tasks. SConU augments conformal uncertainty with significance testing to detect and filter uncertainty outliers before applying the underlying conformal procedure. The key innovation is the construction of two conformal p-values over uncertainty statistics; the first compares a test sample's uncertainty to the calibration distribution, while the second, SConU-Pro, further discounts calibration examples that themselves fail prediction at the target risk level. In the experiments outlined in Section 5, SConU is evaluated in Figure 6a. On MMLU at a target coverage of 90%, SConU achieves the 2$^{\text{nd}}$ lowest APSS out of the eight other methods, though it had a significant overcoverage of $\approx 9.5\%$. A key limitation is that it only approximates conditional coverage, and performance depends on the chosen uncertainty measure and sampling budget.

(b) Logit-free Methods

This section is organised as follows. Section (i) surveys logit-free methods applied to open-ended tasks and Section (ii) presents a comparison between logit-free and logit-based methods.

(i) Open-ended Logit-free Split-conformal Calibration

All split-conformal-based methods outlined in Section 7 require access to LLM token-level logits, logprobs, or model embeddings (see Section 3). This means that black-box models and those with API-only access cannot be conformalised by such methods; many proprietary and frontier LLMs such as Claude Opus, Gemini, and ChatGPT are therefore unable to be conformalised due to their black-box, API-only configuration. Motivated by this limitation, logit-free conformal prediction (LofreeCP) is proposed to apply the split-conformal framework to open-ended QA and close-ended MCQA tasks where the LLM is truly black-box and there is no access to internal probabilities [73]. Such a method should, as with logit-based methods, preserve the coverage guarantee while remaining efficient in terms of prediction set size.

The key idea behind LofreeCP is to derive nonconformity scores from repeated sampled outputs rather than from model logits/logprobs by combining one coarse-grain and two fine-grained uncertainty notions; the first is frequency. For a given prompt x_i if $\hat{y}_i^a$ is the a-th unique response among m samples, its empirical frequency score is $\tilde{F}(\hat{y}_i^a, m) = \frac{\tilde{p}(\hat{y}_i^a)}{m}$, where $\tilde{p}(\hat{y}_i^a)$ is the absolute count of that response in the sample pool. Frequency is motivated as a proxy for uncertainty ranking rather than a literal probability estimate, with empirical analysis showing a positive correlation between response frequency and average true model probability, which suggests that more frequently repeated answers are typically higher-confidence outputs. The authors prove that estimating actual probabilities to high precision by sampling alone is too expensive, finding that achieving 95% confidence with a 1% margin of error would require at least 9,604 samples. As such, LofreeCP provides a ranking signal from a moderate number of samples, rather than attempting accurate probability estimation from the API.

Frequency alone creates a concentration problem; many responses receive the same score because they occur with the same count. To mitigate this, two fine-grained uncertainty measures are added. The prompt-wise fine-grained term is normalised entropy (NE), defined as

$$H\left(x_i \,\middle|\, \left\{\hat{y}_i^j\right\}_{j=1}^{m}\right) = \frac{\sum_{a=1}^{k} \tilde{p}(\hat{y}_i^a) \log \tilde{p}(\hat{y}_i^a)}{\log m}, \tag{7.3}$$

where k is the number of distinct sampled responses among m responses. Intuitively, this measures how dispersed the sampled answers are for a prompt; higher entropy is interpreted as higher uncertainty. This is supported empirically on TriviaQA, which shows that prompts with larger NE are more likely to be questions for which the model fails to sample the correct label.

The response-wise fine-grained term is semantic similarity, defined between a candidate response $\hat{y}_i^a$ and the most frequent response for the same prompt,

$$SS\left(\hat{y}_i^a, P_i^{\text{highest}}\right) = \frac{v(\hat{y}_i^a) \cdot v\left(p_i^{\text{highest}}\right)}{\left\|v(\hat{y}_i^a)\right\| \left\|v\left(p_i^{\text{highest}}\right)\right\|}, \tag{7.4}$$

where $v(x)$ is the vector representation of x, and p_i^{highest} is the response having the highest frequency for the prompt x_i. The intuition is that among two responses with equal frequency, the one semantically closer to the dominant sampled response is likely to be less uncertain. The three uncertainty notions are combined into the nonconformity score $s(x_i, \hat{y}_i^a) = -\tilde{F}(\hat{y}_i^a, m) + \lambda_1 H(\cdot) - \lambda_2 SS(\cdot)$ where $\lambda = (\lambda_1, \lambda_2)$ controls the trade-off between coarse-grained and fine-grained uncertainty. Calibration follows standard split-conformal; the method samples m responses per prompt and computes the true-label nonconformity scores, these calibration scores then compute the empirical quantile $\hat{q}$ corresponding to the desired error level α, and then test-time sampled responses with scores below that threshold are retained. The marginal coverage provided is $\mathbb{P}\{y_{\text{test}} \in C(x_{\text{test}})\} \geq 1 - \alpha$. In the experiments outlined in Section 5, LofreeCP is evaluated in Figure 6b. On TriviaQA at a target coverage of 70%, LofreeCP maintains tight coverage and is ranked 3$^{\text{rd}}$ in terms of the lowest APSS out of the nine other methods evaluated, eight of which are logit-based, which supports the strength of logit-free sampling.

An alternative to LofreeCP is conformal uncertainty or ConU, a logit-free LLM uncertainty method for open-ended text generation tasks [75]. The authors of ConU argue that LofreeCP does not explicitly tie the nonconformity score to the uncertainty condition of correct answers strongly enough. Their central claim is that a more robust conformal criterion can be obtained by defining the calibration nonconformity score directly from the uncertainty of the candidate generation that is both correct and most semantically aligned with the reference answer.

For each prompt x_i, the method samples M candidate generations $\{\hat{y}_i^m\}_{m=1}^M$, then clusters them into K distinct semantics, and defines semantic frequency, $F\left(\hat{y}_i^k\right) = \frac{V_k}{M}$, where V_k is the number of generations in the k-th cluster. This directly reflects the self-consistency intuition, which is similar to LofreeCP; semantics that recur more often across repeated samples are treated as more reliable. Using these semantic clusters, the uncertainty score of each sampled generation $\hat{y}_i^m$ is defined as $U(\hat{y}_i^m) = 1 - \lambda F(\hat{y}_i^m) - (1 - \lambda)\frac{1}{K}\sum_{k=1}^K SS\left(\hat{y}_i^m, \hat{y}_i^k\right) F\left(\hat{y}_i^k\right)$, where SS is a semantic similarity score computed with a DistillRoBERTa-based cross-encoder model, $\lambda \in [0, 1]$ balances direct self-consistency against similarity-weighted agreement with the semantic clusters. The first term penalises low-frequency semantics; the second term rewards a generation for being semantically close to other clusters that themselves have high frequency. This is the main uncertainty measure for individual candidate generations.

A second uncertainty quantity is then defined for the overall query-response process. Let $\hat{y}_i^{\text{mst}}$ denote any generation in the largest semantic cluster, i.e. the model's most self-consistent answer. The uncertainty of the full process is

$$U\left(\{\hat{y}_i^m\}_{m=1}^M \mid x_i\right) = 1 - \lambda \cdot F\left(\hat{y}_i^{\text{mst}}\right) - (1 - \lambda)\frac{1}{K}\sum_{k=1}^K SS\left(\hat{y}_i^{\text{mst}}, \hat{y}_i^k\right) F\left(\hat{y}_i^k\right). \tag{7.5}$$

This quantity is known as ConU. Intuitively, it measures how strongly the dominant semantic is supported by the rest of the sampled semantic landscape. If the most frequent answer is both common and semantically close to other clusters, uncertainty is low; if the sampled semantics are fragmented or contradictory, uncertainty rises. The conformal stage of the method uses the per-generation uncertainty score $U(\hat{y}_i^m)$, not the process-level ConU score, to define the calibration nonconformity score. For each calibration example (x_i, y_i), the sampled generation that is semantically equivalent to the reference answer and has the highest similarity to it is identified. The nonconformity score is then $s(x_i, y_i) = U\left(\arg\max_{\hat{y}_i^j} SS\left(\hat{y}_i^j, y_i\right) E\left(\hat{y}_i^j, y_i\right)\right)$, where $E(\cdot, \cdot)$ is an indicator of semantic equivalence. Thus, the nonconformity score is not just

the uncertainty of some sampled answer, but specifically the uncertainty condition of the sample answer judged correct with respect to the reference. Given the sorted calibration scores $\{s_i\}_{i=1}^{n}$, the split-conformal threshold $\hat{q}$ is chosen as the appropriate empirical quantile corresponding to the target miscoverage level α. At test time, generations with uncertainty below this threshold are included in the prediction set $C(x_{\text{test}}) = \left\{\hat{y}^j_{\text{test}} : s\left(x_{\text{test}}, \hat{y}^j_{\text{test}}\right) \leq \hat{q}\right\}$.

In the experiments outlined in Section 5, ConU is evaluated in Figures 6a, 6b, and 6c. On MMLU at a target coverage of 90%, ConU is ranked 1st in terms of the lowest APSS out of the eight other methods, where one was logit-free and the rest logit-based. On TriviaQA at a target coverage of 90%, ConU is ranked 2nd in terms of the lowest APSS out of the six other methods, outranking the other three logit-free methods. On MedMCQA at a target coverage of 90%, ConU is ranked joint 1st in terms of the lowest APSS out of the four other methods and their variants, outranking the other two logit-free methods. For all three datasets, ConU met the coverage target with a slight overcoverage of $\approx$ 1-2%.

So far, these methods have operated in the unbounded space of open-ended generation, where the main difficulty is in constructing useful prediction sets when the correct answer may lie in the unobserved part of the output space. With current sampling-based conformal methods, if the correct answer has not appeared among a finite number of queried generations, then at high target coverage levels the predictor may be forced into essentially vacuous prediction sets. This becomes acute when the user requests coverage substantially above the few-shot accuracy of the underlying LLM. Therefore, conformal prediction with query oracle (CPQ), a framework that explicitly manages the trade-off between coverage, test-time query budget, and informativeness, is proposed [106]. CPQ interprets this trade-off through the classical missing mass problem in statistics; after sampling a finite number of outputs, how much probability remains on labels that have not yet been seen? This missing mass determines both whether one should keep querying and how much uncertainty should remain in the final prediction set.

A practical detail to note is clustering. Since LLM outputs may differ lexically while conveying the same meaning, CPQ groups sampled generations into semantic equivalence classes, each treated as a single label $y \in \mathcal{Y}$. The frequencies of these clusters are then used for the Good-Turing missing mass estimates. In the experiments outlined in Section 5, the full CPQ variant (P1+P2) is evaluated in Figure 6c. On GSM8K at a target coverage of 70%, CPQ is ranked 2nd in terms of the lowest APSS out of the seven other methods evaluated, where two were logit-free and the rest logit-based. CPQ is also ranked 1st in terms of the tightest coverage out of the five other methods.

So far, logit-free conformal methods have been designed for LLMs and the text modality. TRON, a two-step sampling-based conformal framework [69], is proposed to adapt open-ended logit-free conformal calibration to multimodal tasks, incorporating the visual and audio modalities. It can handle tasks such as visual QA, open-ended video QA (OpenVideoQA), and close-ended multiple-choice video QA (MCVideoQA), but is also versatile and capable of handling text-based tasks such as OpenQA and MCQA with standard unimodal LLMs.

Given a calibration dataset $\{(x_i, y_i)\}_{i=1}^{N}$, for each calibration query x_i, the model can generate M_i sampled responses $\{\hat{y}_i^m\}_{m=1}^{M_i}$, and for the test query, the model generates M_{test} responses $\{\hat{y}^m_{\text{test}}\}_{m=1}^{M_{\text{test}}}$. The overall goal is to construct a prediction set $C(x_{\text{test}})$ such that the correctness miscoverage $\mathbb{P}(y_{\text{test}} \notin C(x_{\text{test}}))$ is controlled by a user-specified risk level.
TRON introduces two user-chosen risk levels:

- α, the risk that the sampled candidate set fails to contain any acceptable answer, and
- β, the risk that the final filtered prediction set fails to contain an acceptable answer given that one was sampled.

The composite risk bound is $\varepsilon = \alpha + \beta - \alpha\beta$, and TRON guarantees $\mathbb{P}(y_{\text{test}} \notin C(x_{\text{test}})) \leq \varepsilon$.

In the experiments outlined in Section 5, we do not perform any evaluation of TRON on video datasets, as this was the only method designed for video QA. However, TRON was shown to generalise to OpenQA tasks with LLMs, and so TRON is evaluated in Figure 6b. On TriviaQA at a target coverage of 90%, TRON achieves the 3rd lowest APSS out of the six other methods and their variants, though it had a slight undercoverage of $\approx$ 1.5%. In terms of limitations, TRON

depends on sampling efficiency; the first stage on works well if acceptable responses are likely to appear a reasonable finite number of samples. If the model rarely generates correct responses, the required $\hat{r}$ could become large, increasing cost and latency. Lastly, TRON depends on semantic equivalence judgements for clustering and reference matching, using NLI-based bidirectional entailment, thereby only as reliable as this evaluator; errors in semantic clustering can negatively affect both the frequency scores and the deduplication-based APSS.

(ii) Performance Comparison of Logit-free and Logit-based Methods

In the review objectives outlined in Section 1(a), we proposed **Hypothesis A**. To recap, we hypothesised that the uncertainty signals taken from conformal methods based on logit-free techniques would be noisier than those taken from logit-based methods. We posit that we would **see a higher coverage error and larger and less informative prediction sets when using logit-free methods**, and that the opposite would be true for logit-based methods, whose uncertainty signal is expected to be a richer and more direct view of a model's confidence.

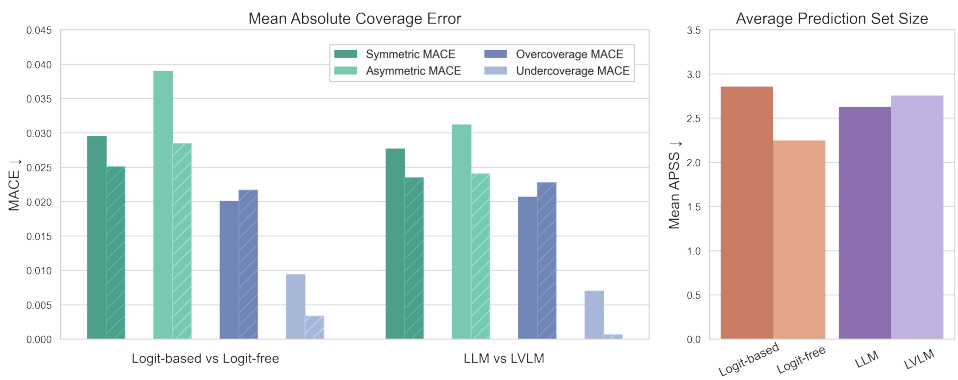

Figure 11: We find that logit-free methods tend to outperform logit-based methods, with lower mean absolute coverage error and prediction set size, but a slightly higher overcoverage error. LVLM methods, compared to LLM methods, also have a lower mean absolute coverage error and very low undercoverage error, but have a higher overcoverage error and prediction set size.

In Section 5, we outlined a controlled performance analysis, from which we conducted additional subgroup studies using a diverse subset of those methods. One such subset was divided based on whether a conformal method used white-box logits, log-probabilities (logprobs), or model embeddings, or whether it used black-box techniques such as reprompting, self-consistency sampling, or entailment scores. We then computed four variants of mean absolute coverage error (MACE) and mean average prediction set size (APSS) as defined in Section 5, and report our results in Figure 11, under 'Logit-based vs Logit-free' and the bars with orange hue.

We find that overall, logit-free conformal methods exhibited a moderately lower symmetric MACE (sMACE), and significantly lower asymmetric MACE (aMACE) and undercoverage error. aMACE penalises failure to meet the nominal coverage more heavily, so this is reasonable. We also observe a non-trivial reduction in APSS for logit-free methods, and find that they tend to be a little more conservative on average, exhibiting a slightly higher overcoverage error than logit-based.

The observations we made did not support our initial hypothesis, and we propose two primary conditions for why this might be the case. There is strong evidence in the literature that logits and their derivatives such as logprobs are miscalibrated. Under in-context learning and particularly in low-shot settings, LLMs have been found to be miscalibrated, with fine-tuning and chain-of-thought (CoT) prompting empirically shown to worsen calibration [215]. Other studies have shown that, in reinforcement learning with human feedback (RLHF)-aligned

LLMs, overconfidence is prevalent, meaning that an LLM's expressed confidence does not match its actual correctness rate [216]. Another study performed an extensive calibration benchmark with 15 chat-tuned LLMs across five benchmark datasets, and found systematic overconfidence; very high maximum softmax probabilities were correct far less often than their probabilities implied [217]. The mismatch was further quantified with ECE, finding substantial calibration error across models. Cross-model analysis found no statistically significant relationship between model accuracy and calibration error for chat LLMs, indicating that stronger chat fine-tuned LLMs were not becoming better calibrated [217].

Overall, the majority of studies show that logits/logprobs are miscalibrated, but findings are limited to close-ended, discrete tasks such as text classification and MCQA, due to calibration being harder to define in open-ended generation settings. Despite this, there have been studies that show empirically that LLMs are also miscalibrated on open-ended tasks. One such study evaluates nine LLMs on open-domain factual QA and finds similar overconfidence and calibration error as previous studies [217], demonstrating that it persists across model families, parameter sizes, and query types [4]. The distribution of tasks in our assessment pool of conformal studies was approximately 60% close-ended tasks such as MCQA and MCVQA, and 40% open-ended tasks such as open-domain QA and mathematical reasoning.

Following the evidence, one would expect the logits used by logit-based conformal methods to also be miscalibrated. Thus, condition one is that **miscalibration may contribute to degrading conformal efficiency** with larger prediction sets due to noisy or unevenly scaled logits, and that **higher undercoverage error could be a symptom of degraded conformal validity**, with overly optimistic or poorly ordered logits enabling conformal thresholds to admit incorrect outputs.

Condition two is that black-box uncertainty measures, such as sampling consistency, semantic diversity, and entailment-based scores may provide better-aligned conformity signals than token-level logits. These signals **may more directly capture semantic correctness or answer stability**, whereas logits and their derivations like log-probabilities reflect local decoding confidence and can be distorted by calibration error and tokenisation effects. As a result of a better aligned conformity signal, logit-free methods, despite being slightly more conservative (as indicated by higher overcoverage error), **may rank candidates more meaningfully, translating to better conformal efficiency**. Their behaviour can be considered more stable, with lower undercoverage indicating that incorrect outputs are identified more consistently than logit-based methods.

(c) Token-level Uncertainty

A central approach to token-level uncertainty in large language models is to interpret the model's next-token distribution as a measure of predictive confidence. Under nucleus (top-p) sampling, this uncertainty is summarised by a prediction set consisting of the smallest collection of tokens whose cumulative probability exceeds p. This induces an appealing probabilistic interpretation; namely, that the set should contain the true next token with probability p, but that interpretation is not guaranteed to hold in practice. This motivates the use of conformal prediction to assess whether token-level prediction sets are calibrated and, when necessary, to recalibrate them.

Conformal nucleus sampling (CNS) is a method proposed to adapt the split-conformal framework to next-token prediction using the APS nonconformity score function (see Section 4) [51]. For each validation example, the conformal score is the minimum cumulative probability mass needed for the LLM's predicted set to include the true next token. The calibrated threshold $\hat{q}$ is then chosen as the relevant quantile of these scores, yielding a conformal top-p decoding rule. This threshold is then calibrated as a function of the entropy of the next-word distribution, as the degree of overconfidence varies strongly across entropy regimes. Experiments across several LLMs show systematic overconfidence, especially at low entropy, and reveal a moderate inverse scaling trend whereby calibration worsens with LLM parameter size. The primary limitation of CNS is that autoregressive language generation violates the exchangeability assumption, and as such, the conformal guarantees are only approximate.

To address this limitation of CNS, non-exchangeable conformal nucleus sampling (NE-CNS) is proposed, which unifies non-exchangeable conformal prediction [43] with nearest-neighbour retrieval in latent decoder space [56]. At each decoding step, the LLM uses the current decoder hidden state to retrieve similar representations from a datastore built from calibration examples, together with their nonconformity scores. Distances to these neighbours are converted into weights, and a weighted conformal quantile is computed to construct a prediction set for the next token. The APS nonconformity score function is utilised (see Section 4). This procedure yields token-level conformal sets that are dynamically calibrated to the local neighbourhood rather than to a global exchangeable calibration pool. NE-CNS has some limitations; first, the theoretical guarantee is weaker than in the standard split-conformal framework because the non-exchangeability penalty depends on a total variation term that is difficult to estimate in practice, and second, nearest-neighbour retrieval introduces extra computational cost at inference time.

Table 3: Performance of split conformal prediction methods and some conformal risk control methods. ECR and SSC are reported as percentages where closer to the target coverage is better, APSS and NSR are reported as rates where lower values are better; for definitions see Section 5, Table 2. Where possible the error rate used is $\alpha = 0.10$ (nominal coverage $1 - \alpha = 0.90$) but this may not be true for all cases, please refer to the performance notes.

Ref	Conformal Method	ECR	APSS↓	SCC	NSR↓	Performance Notes
[50]	LAC	92.3	3.28	60.0		Mean across 16 subjects in MMLU. SSC: min over prediction set size bins of stratified ECR.
[53]	CLM	70.1	1.12	68.3		Using `max` scoring on TriviaQA. ECR/SSC: fixed $\alpha = 0.30$ (nominal $1 - \alpha = 0.70$) from [73]. APSS: normalised set size at $\alpha = 0.30$ (nominal $1 - \alpha = 0.70$) $\times\ k_{\max}$; $k_{\max} = 20$.
[80]	LAC+APS	91.4	2.86			Mean of per-LLM means (5 tasks) across 11 LLMs. All per-LLM means are average of LAC and APS values.
[94]	LAC	88.0	3.01			Mean across 8 LVLMs $\times$ 2 datasets.
[145]	COPU	90.4	1.80			Mean across 6 LLMs $\times$ 4 datasets.
[100]	CROQ	95.2	1.94			Mean across 3 LLMs $\times$ 3 datasets at fixed $\alpha = 0.30$ (nominal $1 - \alpha = 0.70$), option count = 4 (standard MCQA).
[82]	CPL	91.2	2.58		0.644	Mean across 5 datasets. NSR: MMLU at fixed $\alpha = 0.05$ (nominal $1 - \alpha = 0.95$) from [146].
[60]	TRAQ	95.6	29.25			ECR: mean of per-LLM means (3 datasets), APSS: semantic counts; mean of per-LLM means (GPT-3.5 over 4 datasets; LLaMA-2 over 3).
[60]	TRAQ-P	98.1	39.12			Same as above.
[52]	KnowNo	92.0	2.59		0.880	From [83].
[101]	GPS	96.6	23.98			Using GPS HS. ECR: mean across 13 model-dataset pairs (5 datasets, 5 LLMs); APSS: mean over finite entries only ($n = 10$; 3 entries report set size as '-').
[106]	CPQ	89.0	1.13			Using full CPQ (P1+P2). Mean across 3 datasets.
[73]	LofreeCP	70.3	1.11	76.6		Mean across 2 datasets at $\alpha = 0.30$ (nominal $1 - \alpha = 0.70$).
[85]	S-ATLAS	94.6			0.044	
[88]	CAP	92.9	2.12			Mean of per-LLM precomputed means (5 datasets) across 3 LLMs.
[88]	CAP	92.9	2.14			Mean of per-LVLM precomputed means (5 datasets) across 3 LVLMs.
[86]	TECP	97.3	8.99			Mean across 6 LLMs $\times$ 2 datasets.
[107]	C-IP	89.0				Mean across 3 specialities in MediQ. ECR: final iteration $k = 8$.
[83]	I-CP	98.5	2.27		0.750	
[75]	ConU	92.0	2.53			Mean across 7 LLMs $\times$ 4 datasets.
[51]	CNS	99.4	0.93	64.1		Task-balanced mean across 2 tasks; mean machine translation (2 directions $\times$ 2 models) and LM (2 models).
[56]	NE-CNS	90.6	0.21	74.4		Same as above.
[108]	SConU	90.5				Mean across 2 LLMs $\times$ 2 subjects in MMLU-Pro.
[108]	SConU-Pro	91	1.97	68.9		ECR/APSS: mean over off-diagonal entries of subject-transfer matrix at $\alpha = \alpha_\ell = 0.2723$ (nominal $1 - \alpha \approx 0.73$), where α_ℓ is minimum feasible risk level on MMLU-Pro. SSC: mean SConU at `split_ratio`$= 0.5$, $\alpha = \alpha_\ell = 0.3342$ (nominal $1 - \alpha \approx 0.66$) on MedMCQA.
[90]	ConfTS-APS	90.0	2.05			Mean across 2 datasets. ECR not tabulated but maintains desired marginal coverage.
[90]	ConfTS-RAPS	90.0	2.03			Same as above.

Continued on next page

Table 3 continued from previous page

Ref	Conformal Method	ECR	APSS↓	SCC	NSR↓	Performance Notes
[95]	SafePath	87.0	1.00			ECR $= 1 - (\alpha) +$ DTC.
[120]	HERACLES	92.4			0.125	Mean across 3 LLMs. Note that in the general (non-experimental) setting, help rate can include physical infeasibility.
[96]	Conformal NL2LTL	95.3			3.410	Fixed $\alpha = 0.05$ (nominal $1 - \alpha = 0.95$).
[66]	SCP-SE	76.0	5.00			Fixed $\alpha = 0.30$ (nominal $1 - \alpha = 0.70$).
[146]	SOCOP	95.0	2.48		0.587	Fixed $\alpha = 0.05$ (nominal $1 - \alpha = 0.95$).
[68]	Conformal Reasoning	89.2	2.88		0.656	NSR $= 100 -$ %answered. APSS $= (1 -$ specificity$) \times \lvert Y \rvert$, where $\lvert Y \rvert = 4$ for MediQ/EQA, $\lvert Y \rvert = 1854$ for 20Q.
[114]	SCP-ST	90.7	4.60			Mean across 7 LLMs $\times$ 2 datasets.
[77]	SCP-SR	85.5	35.40			Mean across 2 datasets.
[97]	SCP-SC logit	89.8	3.60			Mean across 4 LLMs on MedMCQA.
[97]	SCP-SC logit-free	95.5	3.82			ECR: mean across 4 LLMs $\times$ 2 datasets + 4 subjects in MMLU. APSS: mean across 4 LLMs $\times$ 3 datasets.
[98]	SCOPE-Gen logit	78.0	1.20			ECR: on BBH Date Understanding from [106]. APSS: using `sum` scoring with quality score. Fixed $\alpha = 0.30$ (nominal $1 - \alpha = 0.70$).
[98]	SCOPE-Gen logit-free	78.0	1.26			Same as above, but APSS: using `count` scoring without quality score.
[149]	CQR	95.4	2.64			Mean across 3 LLMs $\times$ 5 datasets (where SummEval is mean of 4 dimensions) after boundary adjustment. Note: whilst APSS and AIW are conceptually analogous, APSS is discrete whereas AIW is continuous; we consider AIW equivalent to APSS as an efficiency metric in this case.
[149]	Asym CQR	96.8	2.78			Same as above.
[149]	CHR	87.1	1.68			Same as above.
[149]	LVD	93.9	2.19			Same as above.
[149]	Boosted CQR	92.9	2.14			Same as above.
[149]	Boosted LCP	92.6	2.19			Same as above.
[149]	R2CCP	91.0	1.86			Same as above.
[149]	OrdinalAPS	71.6	1.72			Same as above.
[149]	OrdinalRC	72.8	1.79			Same as above.
[69]	TRON	86.8	1.80	84.0		ECR: mean across 2 LLMs $\times$ 4 datasets. APSS: mean across 8 LLMs $\times$ 4 datasets. SSC: mean across 3 set sizes $\times$ 4 datasets. Fixed user-specified risk levels α (sampling) and β (identification), where $\varepsilon = \alpha + \beta - \alpha\beta$; $\alpha = \beta = 0.1$; $\varepsilon = 0.19$ (nominal $1 - \varepsilon = 0.81$), or $(1 - \alpha)(1 - \beta) = 0.81$.
[69]	TRON LLM	86.8	2.63			ECR: mean across 6 LLMs $\times$ 3 datasets. APSS: mean across 2 LLMs on MMLU with $M = 20$. Error level defined same as above.
[132]	DS-CP	92.0	2.30			Mean across 16 LLMs.
[74]	SAPS	70.0	1.55	49.4		On TriviaQA at fixed $\alpha = 0.30$ (nominal $1 - \alpha = 0.70$).
[57]	DCBS	82.3	1.76			On Integer Addition at fixed $\alpha = 0.05$ (nominal $1 - \alpha = 0.95$).
[104]	PC-SGG	78.5	413.36			For triplet prediction sets with LVLM plausibility at $\varepsilon = \alpha_o + \alpha_r - \alpha_o\alpha_r$; $\alpha_o = 0.05$ (object), $\alpha_r = 0.1$ (predicate); $\varepsilon = 0.145$ (nominal $1 - \varepsilon = 0.855 \approx 0.86$), or $(1 - \alpha_o)(1 - \alpha_r) = 0.855 \approx 0.86$.
[135]	AR-NECP	90.3	3.73			Mean across 3 LLMs on TriviaQA with WHOLE shift using reweight balancing.
[124]	LAC+APS	93.3	2.65			Mean of per-LVLM precomputed means (5 datasets) across 10 LVLMs. All precomputed means include average of LAC and APS values.
[158]	ICAD	96.2				Mean of per-LLM means (2 datasets) across 2 LLMs for $K = 3$ layers $\in \{7, 15, 22\}$.
[159]	CoFineLLM	99.0	1.15		0.133	For in-distribution scenarios with $\lambda = 0.1$.
[133]	SAFER	91.7	7.20			Mean across 5 LLMs on TriviaQA at fixed $\varepsilon = \alpha + \beta - \alpha\beta$; $\alpha = \beta = 0.05$; $\varepsilon = 0.0975$ (nominal $1 - \varepsilon = 0.9025 \approx 0.90$), or $(1 - \alpha)(1 - \beta) = 0.9025 \approx 0.90$. In $\varepsilon = \alpha + \beta - \alpha\beta$, α is sampling risk level, β is filtering risk level, and ε is overall error level.
[133]	SAFER	91.7	7.20			Mean across 5 LLMs on CoQA at $\varepsilon = \alpha + \beta - \alpha\beta$; $\alpha = 0.25, \beta = 0.05$; $\varepsilon = 0.2875$ (nominal $1 - \varepsilon = 0.7125 \approx 0.70$), or $(1 - \alpha)(1 - \beta) = 0.7125 \approx 0.70$. Error level defined same as above.
[71]	UAQ	73.0	3.00			On WebQuestionsSP at fixed $\alpha = 0.30$ (nominal $1 - \alpha = 0.70$).
[160]	SarRec		7.33			Mean across 3 datasets.
[113]	CoAlign	92.0				Whilst PSR is reported, we cannot directly recover APSS as average total candidates is omitted for commercial privacy.

Continued on next page

Table 3 continued from previous page

Ref	Conformal Method	ECR	APSS↓	SCC	NSR↓	Performance Notes
[162]	ConMIL	95.1	2.21		0.734	Mean across 2 datasets, NSR mean across 2 LLMs × 2 datasets, at fixed $\alpha = 0.05$ (nominal $1 - \alpha = 0.95$). NSR = $\frac{\text{uncertain samples}}{N}$. ECR/APSS found in supplementary materials.
[93]	Conformal Tree	89.0	2.96			Mean across 6 classes on UCI Erythemato-Squamous Disease.
[118]	SCP-PE	98.3	4.38			Mean across 6 LLMs × 2 datasets.
[126]	LAC	90.4	4.42			Mean across 16 LVLMs × 6 datasets.
[126]	MS	93.6	4.58			Same as above.
[126]	APS	90.8	3.96			Same as above.
[127]	BB-UCP	88.0				Mean across 6 datasets for cross-query task.
[141]	PA-SCP	90.4	1.92			Mean across 5 tasks for fully-reworded.
[141]	PA-QCCP	90.5	1.92			Same as above.
[142]	CR^2		1.28			Fixed tolerance $\alpha = 0.10$.
[164]	CFC	69.6	1.12			Mean across 2 datasets using the FULL variants at fixed $\alpha = 0.30$ (nominal $1 - \alpha = 0.70$).
[164]	CFC-PAC	70.5	1.24			Same as above.

8. Conformal Selective Prediction

This section covers conformal selective prediction methods illustrated in Figure 2, particularly those based on the conformal selection framework [34,35]. The overall section is outlined as follows. First we survey conformal selection methods for LLMs, then the performance of all FDR-controlling methods is reported in Table 4, and lastly Figure 12 demonstrates the conformal selective procedure with an example that showcases selection, deferral, and abstention.

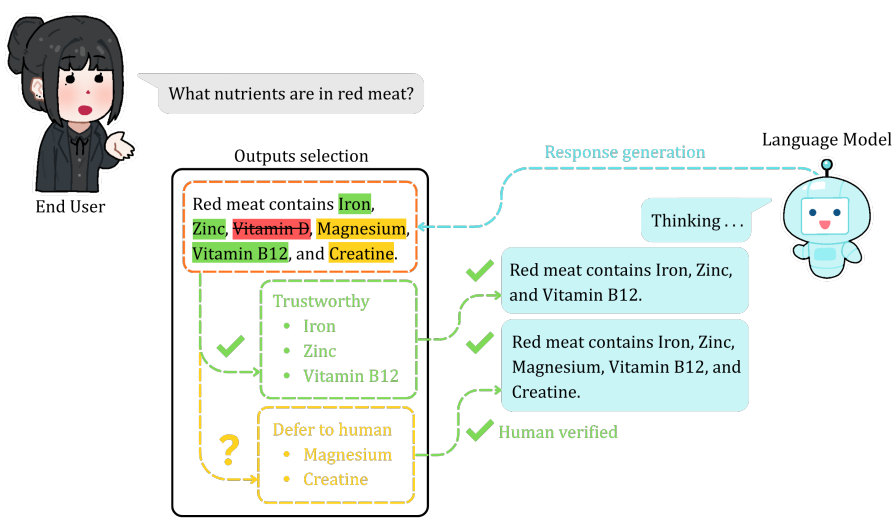

Figure 12: **Conformal Selective Prediction** enables the automatic, rule-based selection of trustworthy generated outputs. Less certain outputs can be deferred (for example, to a human for review), while untrustworthy outputs are rejected through abstention. In this example, trustworthy outputs are highlighted green, uncertain outputs which are deferred are highlighted yellow, and untrustworthy outputs which are rejected are highlighted red. The green check mark indicates that the answer is correct and that the deferred outputs were validated as trustworthy.

In high-stakes settings, such as open question-answering (OpenQA) and radiology report generation (RRG), knowing an LLM is "good on average" is not enough, and one needs a principled way to determine which specific outputs can be trusted. Existing conformal methods

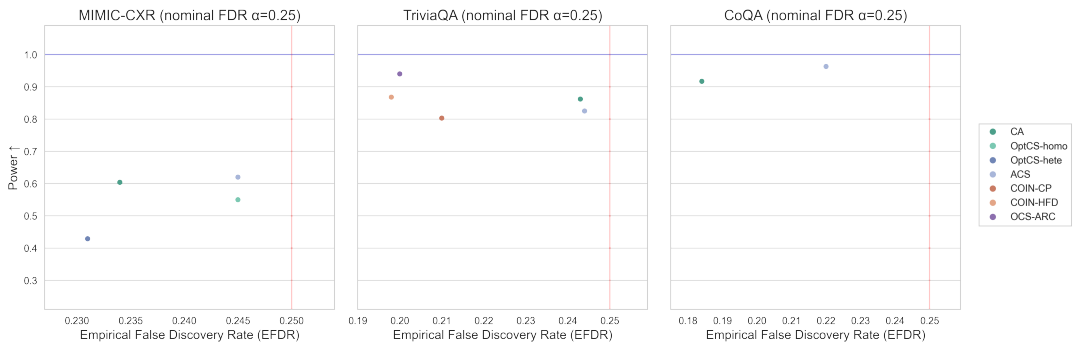

Figure 13: Scatterplots for conformalised selection-based methods on MIMIC-CXR, TriviaQA, and CoQA. Metrics are defined in Section 5. The x-axis is the empirical FDR, and the red line indicates the nominal FDR level which all methods must remain below to be valid. The y-axis is the power; higher values are better, and the blue line indicates maximum power for a perfect method.

such as those based on the split-conformal framework yield prediction sets, but coverage of prediction sets does not directly guarantee that the selected units are reliable for deployment. Other conformal methods either guarantee that at least one output in a set is aligned, which leaves manual disambiguation to the end user, or modify the generated output to make it safer, which may reduce informativeness.

Conformal alignment therefore proposes a different target; rather than altering the generated output conservatively or returning multiple candidates, it seeks to certify which already-generated outputs are trustworthy and abstain from the rest [81]. This translates to a safer LLM deployment pipeline as selected units can be used automatically, while non-selected units can be deferred to humans. Let $f : \mathcal{X} \to \mathcal{Y}$ be a fixed pre-trained LLM that maps an input x to an output $f(x)$. Access to a holdout dataset $D = (x_i, e_i)_{i=1}^n$ is assumed, where $x_i \in \mathcal{X}$ is the input and $e_i \in \mathcal{E}$ is reference information that can be used to assess whether the output is aligned. An alignment function $A : \mathcal{Y} \times \mathcal{E} \to \mathbb{R}$ assigns a true alignment score to the generated output $f(x)$ relative to the reference E. For example, in the radiology setting, this score measures similarity between a generated report and a human expert report. Given a test set $D_{\text{test}} = \{x_{n+j}\}_{j=1}^m$, the goal is to select a subset $S \subseteq \{1, ..., m\}$ of units such that most selected outputs exceed a user-specified alignment threshold c. The error metric is the false discovery rate (FDR)

$$\text{FDR} = \mathbb{E}\left[\frac{\sum_{j=1}^m \mathbf{1}(A_{n+j} \le c,\ j \in S)}{\max(|S|, 1)}\right] \le \alpha, \tag{8.1}$$

for a prescribed level $\alpha \in (0, 1)$. Thus, the guarantee is that only a α-fraction of selected outputs S are misaligned on average. The corresponding notion of utility (Power) is

$$\text{Power} = \mathbb{E}\left[\frac{\sum_{j=1}^m \mathbf{1}(A_{n+j} > c,\ j \in S)}{\max\left(\sum_{j=1}^m \mathbf{1}(A_{n+j} > c),\ 1\right)}\right]. \tag{8.2}$$

Therefore, the objective is not merely to control FDR, but to do so while selecting as many truly aligned units as possible. A conformal p-value for the test unit x_{n+j} is defined as

$$p_j = \frac{1 + \sum_{i \in D_{\text{cal}}} \mathbf{1}(\hat{A}_i \le c,\ \hat{A}_i \ge \hat{A}_{n+j})}{|D_{\text{cal}}| + 1}, \quad \hat{A}_i = g(x_i). \tag{8.3}$$

Intuitively, p_j is small when the test unit's predicted score $\hat{A}_{n+j}$ is larger than most calibration units that are actually misaligned. Thus, a small p-value provides evidence that a unit's output is trustworthy. Once p-values are computed, the Benjamini-Hochberg (BH) procedure is applied. Then, the selected set is $S = \{j = 1, ..., m : p_j \le \alpha k^*/m\}$. Under exchangeability and with a

no-ties condition on predicted alignment scores, FDR $\leq \alpha$ is guaranteed. This is finite-sample and distribution-free, and is not on per-unit correctness, but on the average proportion of false discoveries among selected units. Conformal alignment is itself agnostic to the alignment criterion; it only needs a reference-based scoring rule and a predictor for alignment likelihood. In addition, an asymptotic characterisation of power is provided, and used to argue that power depends on two factors: the intrinsic quality of the LLM and the informativeness of the predictor g in distinguishing aligned from non-aligned units. For the OpenQA task, the alignment score is binary $A_i \in \{0, 1\}$ and defined by ROUGE-L similarity, and for the RRG task, defined using CheXbert clinical-observation vectors. Several feature families are used to train g, including `Self_Eval` based on the LLM's own stated confidence, `Lexical_Sim`, `Num_Sets`, and `SE`, which summarise uncertainty across 20 sampled generations, and `EigV`, `Deg`, and `Ecc`, which are features derived from graph-based similarity structures over sampled outputs. Graph-based features are often more informative for alignment prediction than self-evaluated confidence.

In the experiments outlined in Section 5, conformal alignment (denoted CA) is evaluated in Figure 13. On TriviaQA with a nominal FDR of 0.25, CA achieved the 3$^{\text{rd}}$ highest power out of the other three methods and their variants whilst satisfying the FDR constraint. On MIMIC-CXR with a nominal FDR of 0.25, CA achieved the 2$^{\text{nd}}$ highest power out of the other two methods and their variants whilst satisfying the FDR constraint. On CoQA with a nominal FDR of 0.25, CA achieved a high power whilst satisfying the FDR constraint.

Potentially the biggest limitation of conformal alignment (and the conformalised selection framework in general) is the rigidity of the framework; the full analysis plan, including the train/calibration split, alignment predictor, conformity score, and ranking of candidates is fixed and must be chosen before inspecting the calibration and test data. This is unrealistic in deployment settings, where practitioners may want to adjust these parameters after seeing partial information. For example, they may want to reallocate labelled data to improve model fitting, change an underfit alignment predictor, prioritise diversity after observing the current selected pool, or incorporate newly revealed labels.

Conformal selection methods cannot do this without risking invalidation of error guarantees; thus, adaptive conformal selection (ACS), an interactive extension of conformal selection that preserves finite-sample FDR control under adaptive data analysis, is proposed [109]. Their approach is based on a testing perspective: it sequentially screens the samples from the least interesting to the most interesting until a stopping criterion is met, and then declares the unscreened test samples as interesting. The setup involves n labelled data $\{z_i = (x_i, y_i)\}_{i=1}^n$, m unlabelled test data $\{x_{n+j}\}_{j=1}^m$, and null hypotheses $\mathcal{H}_{0,j} : y_{n+j}$ is uninteresting.

Initially, ACS trains a classifier $\hat{f}(x)$ on the training set and ranks the calibration and test samples in ascending order of $\hat{f}(x_i)$, from least to most promising. Let k represent the size of the training set (the size of the calibration set is $n - k$) and π be a permutation of $\{k + 1, ..., n + m\}$ such that $\hat{f}(x_{\pi(k+1)}) \leq ... \leq \hat{f}(x_{\pi(n+m)})$. ACS screens samples sequentially in the order defined by π; at step $l > k$, the next sample $\pi(l + 1)$ may be chosen adaptively, but only using information measurable with respect to the filtration

$$F_l = \sigma\Big(O_l,\ |N_l^-|,\ |P_l|,\ \{(\tilde{Z}_i, A_i)\}_{i \in O_l \cup N_k^+},\ \{x_i\}_{i \in U_l}\Big), \tag{8.4}$$

where $|S|$ represents the cardinality of a set S, O_l contains already screened samples, N_l^+ and N_l^- are the sets of non-null and null unscreened labelled samples, respectively, P_l is the set of unscreened test samples, A_i is the membership indicator whether z_i is a labelled sample and U_l is the set of all the unscreened samples. Note that $\tilde{Z}_i = z_i$ if $A_i = 0$ and $\tilde{Z}_i = x_i$ otherwise.

The procedure stops when the false discovery proportion (FDP) estimate at step $l \geq k$

$$\widehat{\mathrm{FDP}}(l) = \frac{m}{n - k + 1} \cdot \frac{1 + |N_l^-|}{\max(|P_l|, 1)}, \tag{8.5}$$

falls below the target level α, and then selects all remaining unscreened test samples P_l. Otherwise, it screens another sample according to the current adaptive ordering rule. The

final selection set is $S_{\text{ACS}} = P_T$, where $T = \inf\{l \geq k : \widehat{\text{FDP}}(l) \leq \alpha\}$. Under exchangeability, ACS controls the FDP at level α. The proof relies on a supermartingale argument. Specifically, $M_l = \frac{|P_l \cap \mathcal{H}_0|}{1+|N_l^-|}$ is shown to be a supermartingale with respect to a filtration slightly richer than $\{\mathcal{F}_l\}_{l \geq k}$ where $\mathcal{H}_0$ denotes the set of true null hypotheses, i.e. "uninteresting" samples. Since T is a stopping time, the optional stopping theorem yields the FDR guarantee.

In the experiments outlined in Section 5, ACS is evaluated in Figure 13. On both MIMIC-CXR and CoQA with a nominal FDR of 0.25, ACS achieved the highest power out of all methods whilst satisfying the FDR constraint. On TriviaQA with a nominal FDR of 0.25, ACS achieved the 2$^{\text{nd}}$ lowest power out of the other three methods and variants, though it satisfied the FDR constraint.

Table 4: Performance of conformal selection methods. EFDR and power are reported as rates, where EFDR $\leq$ target FDR is better, and power $= 1$ is best; for metric definitions see Section 5, Table 2. The target FDR was $\alpha = 0.25$ for all results reported in this table.

Ref	Conformal Method	EFDR	Power ↑	Performance Notes
[81]	Conformal alignment	0.214	0.890	Mean across 2 LLMs × 2 datasets using $\lvert D\rvert = 2000$ with logistic regression as alignment predictor.
[81]	Conformal alignment	0.234	0.604	On MIMIC-CXR with the same as above.
[102]	COIN-CP	0.224	0.844	Using Clopper-Pearson upper confidence bound. EFDR: mean across 5 LLMs × 2 datasets. Power: mean across 4 LLMs × 2 datasets.
[102]	COIN-HFD	0.203	0.808	Using Hoeffding-style upper confidence bound with same as above.
[109]	ACS	0.232	0.894	Mean across 2 LLMs × 2 datasets using ACS with adaptive model selection and $N = 1000$.
[109]	ACS	0.245	0.620	On MIMIC-CXR with the same as above.
[76]	OptCS-Homo	0.245	0.550	Using OptCS-Full-MSel with homogeneous pruning in setup 2 (stable model selection under many comparable candidates) on MIMIC-CXR.
[76]	OptCS-Hete	0.231	0.429	Using OptCS-Full-MSel with heterogeneous pruning with same as above.
[84]	SGenSemi	0.090	0.502	Mean across 2 LLMs with $\lvert Z_U\rvert = 10K$. We consider FDR-E as a proxy to EFDR and efficiency as a proxy to power.
[115]	OCS-ARC	0.200	0.940	Averaged over 100 runs, reported at timestep = 200 (online setting) on TriviaQA. EFDR is averaged FDP.

9. Conformal Abstention

This section covers conformal abstention methods illustrated in Figure 2. Figure 14 is an example LLM interaction where conformal abstention would be appropriate.

Standard split-conformal methods enforce a single global risk level, which imposes the same coverage-informativeness trade-off on every input. This static setup does not always align with the highly variable uncertainty patterns of LLMs, where some instances warrant confident point predictions while others should produce prediction sets or fully abstain from generating an answer. Conformalised abstention policy (CAP) combines the split-conformal framework with reinforcement learning to learn an instance-conditional abstention policy [89]. Rather than fixing a single α, CAP learns per-input risk parameters (α, β) that govern a dual-threshold conformal mechanism. This lets the model choose among three actions; return a point prediction, return a prediction set, or abstain. The policy is optimised with REINFORCE against a utility-driven reward that balances accuracy, set size, abstention cost, coverage, and exploration. In the experiments outlined in Section 5, CAP is evaluated in Figure 6a. On MMLU at a target coverage of 90%, CAP ranked 7$^{\text{th}}$ in terms of lowest APSS out of the eight other methods, and had a slight overcoverage of $\approx 2\%$. A limitation is that it introduces additional optimisation complexity.

Prior conformal abstention methods for open-ended generation typically assume that, for every question, repeated finite sampling will eventually produce at least one admissible or

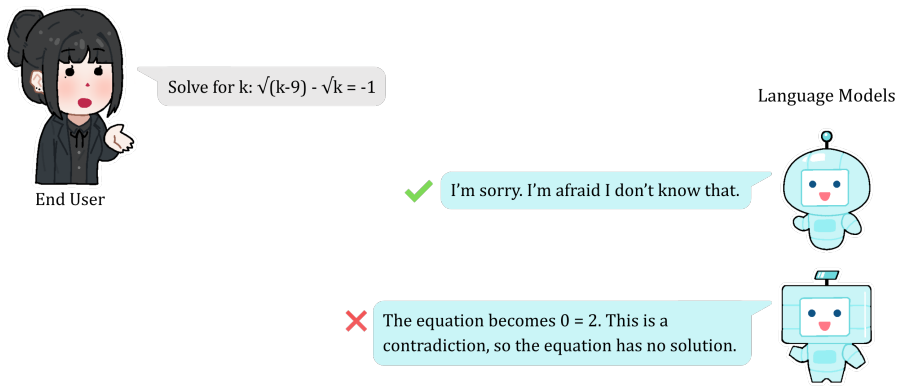

Figure 14: **Conformal Abstention** addresses settings in which the language model cannot provide a reliable answer, such as when the context is ambiguous or the task domain falls outside of the model's parametric knowledge. In such cases, the model abstains from responding rather than generating a plausible but non-factual answer (a hallucination). The green check mark indicates that the answer is correct, and the red cross indicates that it is incorrect.

correct answer. This is unrealistic in open-ended settings, where the output space is unbounded and some questions may simply be beyond a deployed LLM's capabilities under the allowed sample budget. Ignoring this feasibility means calibrated sampling budgets can fail to control the intended test-time coverage risk. In addition, when a candidate set does contain an admissible answer, it often also contains hallucinated, irrelevant, or low-quality distractors; leaving the end user to disambiguate between the outputs, which is inconvenient and unhelpful.

SAFER, a two-stage framework combining abstention-aware sampling budget calibration and conformalised filtering, is proposed to jointly solve both issues [133]. Given calibration data $D_{cal} = \{(x_i, y_i)\}_{i=1}^{N}$ and a deployed LLM $F : \mathcal{X} \to \mathcal{Y}$, for each question $x_i \in \mathcal{X}$, the LLM can sample up to M candidate answers $\{\hat{y}_i^j\}_{j=1}^{M}$. An answer $\hat{y}_i^j$ is considered admissible if it satisfies $A(\hat{y}_i^j, y_i) \geq \lambda$, where $A : \mathcal{Y} \times \mathcal{Y} \to [0, 1]$ is a task-specific relevance function and λ is a user-specified correctness threshold. Several admissibility criteria, including sentence similarity, Rouge-L, bidirectional entailment, and LLM-based semantic evaluation, may be considered.
SAFER introduces two user-chosen risk levels:

- α, the maximum allowable miscoverage rate of the sampling set, and
- β, the maximum allowable miscoverage rate introduced by the filtering step, conditional on the sampling set containing an admissible answer.

In stage one, a maximum sampling cap M is fixed. For each calibration question and each candidate sampling budget $b \leq M$, SAFER counts how many calibration points fail to produce any admissible answer within the first b samples,

$$\hat{m}_{\text{cal}}(b) = \sum_{i=1}^{N} \mathbf{1}(\forall \hat{y} \in \{\hat{y}_i^j\}_{j=1}^{b},\ A(\hat{y}, y_i) < \lambda). \tag{9.1}$$

This yields the empirical miscoverage rate $\hat{r}_{\text{cal}}(b) = \frac{\hat{m}_{\text{cal}}(b)}{N}$. Rather than treating $\hat{r}_{\text{cal}}(b)$ itself as the risk estimate, the Clopper-Pearson exact method is applied to derive an upper confidence bound on the true miscoverage rate δ,

$$\hat{R}^{+}(b) = \sup\left\{R : \mathbb{P}\big(\text{Binom}(N, R) \leq \lceil N\hat{r}_{\text{cal}}(b)\rceil\big) \geq \delta\right\}. \tag{9.2}$$

If no such $\hat{b}$ exists even at M, then SAFER abstains from answering. The calibrated minimum statistically valid sampling budget is $\hat{b} = \inf\left\{s \in \{1, \ldots, M\} : \hat{R}^{+}(b) \leq \alpha\right\}$.

In stage two, SAFER filters the $\hat{b}$-sized candidate set using a calibrated uncertainty threshold. First, it forms a subset of calibration examples for which admissible answers are actually obtainable under budget $\hat{b}$, $D_{\text{cal}}(\hat{b}) = \left\{(x_i, y_i) \in D_{\text{cal}} : \exists \hat{y} \in \{\hat{y}_i^j\}_{j=1}^{\hat{b}} \text{ such that } A(\hat{y}, y_i) \geq \lambda\right\}$. Therefore, only examples for which the sampling stage succeeds are used to calibrate the filtering stage. For a threshold t, SAFER defines the filtered set for each calibration sample (x_i, y_i) in $D_{\text{cal}}(\hat{b})$ as $C_t(x_i) = \left\{\hat{y} \in \{\hat{y}_i^j\}_{j=1}^{\hat{b}} : s(x_i, \hat{y}) \leq t\right\}$, where $s(x_i, \hat{y})$ is an uncertainty score calculated as the accumulated toke-wise entropy over the generated sentence $\hat{y}$. Using conformal risk control, the loss for calibration point (x_i, y_i) at threshold t is $\mathcal{L}_i(t) = \mathbf{1}(\forall \hat{y} \in C_t(x_i),\ A(\hat{y}, y_i) < \lambda)$. The uncertainty threshold t is then determined as

$$\hat{t} = \inf\left\{t : \frac{N'\,\mathcal{L}_{N'}(t) + 1}{N' + 1} \leq \beta\right\}, \tag{9.3}$$

where N' is the size of $D_{\text{cal}}(\hat{b})$, and $\mathcal{L}_{N'}(t) = \frac{1}{N'}\sum_{i=1}^{N'} \mathcal{L}_i(t)$ is the average loss over all samples in the calibration subset. Combining the guarantees from both stages yields an overall high-probability guarantee: with confidence at least $1 - \delta$, the probability that none of the retained answers are admissible is bounded by $\alpha + \beta - \alpha\beta$, where δ is the significance level (i.e., 0.05).

SAFER has some limitations. It relies on the assumption of exchangeability between the calibration and test data, the filtering stage still depends on the quality of the chosen uncertainty heuristic $U(\cdot)$, and the framework is computationally heavier due to repeated sampling up to M times per input and calibration over multiple candidate budgets s, which could increase LLM inference costs. In the experiments outlined in Section 5, SAFER is evaluated in Figure 6b. On TriviaQA at a target coverage of 90%, SAFER achieves the joint 6$^{\text{th}}$ lowest APSS out of the seven other methods and their variants, with a slight overcoverage of $\approx 1.7\%$.

10. Conformal Factuality

This section covers conformal factuality methods illustrated in Figure 2. The overall section is outlined as follows. Section 10(a) examines the principal marginal conformal factuality methods for LLMs, Section 10(b) group-conditional conformal factuality, Section 10(c) distribution-shift-aware conformal factuality, and Section 10(d) conformal factuality with retrieval-augmented generation (RAG). The performance of all methods is reported in Table 5, and Figure 15 demonstrates the conformal factuality procedure for biography generation.

(a) Marginal Conformal Factuality

Conformal factuality is a principled approach to assessing the factuality of LLMs in generative settings such as open-ended QA. The output space in such settings is enormous, with many semantically equivalent responses. In some cases, the prediction set may be intractably large, and in others, may not even contain the correct answer to the question. Rather than an explicit set of candidate completions, the objective for conformal factuality is to construct a prediction set implicitly as the set of all statements that entail a language model's output, where this entailment set satisfies a conformal guarantee. In this way, one may estimate the factuality of claims in long-form answers to open-ended questions, filtering out unsubstantiated claims, while guaranteeing at a user-specified error level a that the retained claims are factual.

Conformal factuality [65] is proposed. Let input $x \in \mathcal{X}$ be mapped by an LLM L to an output $y^* \in \mathcal{Y}$. Correctness is defined relative to a reference answer or knowledge source $y \in \mathcal{Y}$ using an entailment relation $y \Rightarrow y^*$. The target guarantee is given by $\mathbb{P}\big(y^* \text{ is factual and correct}\big) \geq 1 - \alpha$, for the user-chosen $\alpha \in (0, 1)$. The entailment set is defined as $E(y) := \{\, y' \in \mathcal{Y} : y' \Rightarrow y^* \,\}$ so that correctness of y^* is equivalent to the event $y \in E(y^*)$. The split-conformal framework therefore becomes the mechanism by which the model selects how much it should hedge from its original answer to remain factual overall. This is done by first introducing a family of hedging functions

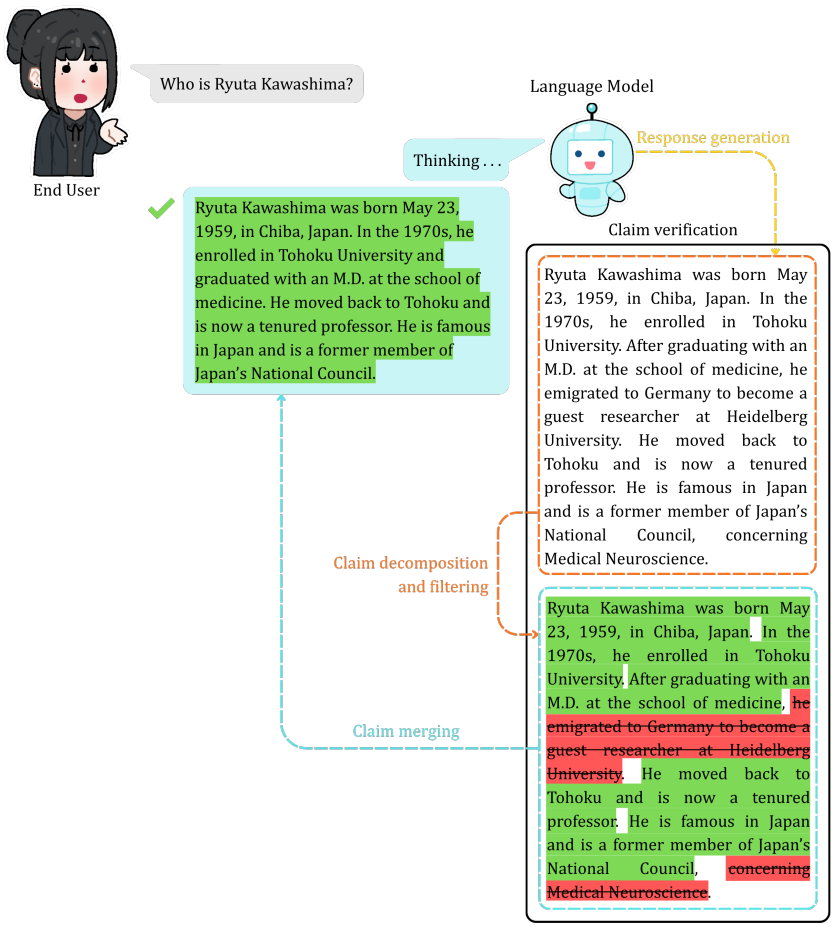

Figure 15: **Conformal Factuality** promotes reliable LLM-generated responses by decomposing long-form text into atomic sub-claims, filtering out non-factual claims, and merging the retained claims back into a coherent response for the end user. In this example, factual and retained claims are highlighted green, non-factual claims which are rejected are highlighted red. The green check mark indicates that the answer is correct.

F_t, indexed by threshold $t \in T \subseteq \mathbb{R}$, where increasing t makes the output less specific by removing uncertain content. As such, the level of detail is diluted to ensure correctness.

A strictly safe conformal score is defined as $r(x, y) = \inf\{t : \forall j \geq t,\ y \in E(F_j(x))\}$. Using the split-conformal framework, the empirical quantile $\hat{q}_\alpha$ of calibration scores $\{(x_i, y_i)\}_{i=1}^m$ is chosen and outputs $F_{\hat{q}_\alpha}(x, L(x))$, yielding fine-sample marginal guarantee, $\mathbb{P}\big(y_{n+1} \in E(F_{\hat{q}_\alpha}(x_{m+1}))\big) \geq 1 - \alpha$, under an exchangeability assumption over calibration and test samples. The standard conformal guarantee is now expressed as a guarantee on the correctness of model-generated text.

In the experiments outlined in Section 5, conformal factuality and partial conformal factuality are denoted CF and pCF, respectively. In Figure 16a, CF is evaluated on MedLFQA for both marginal and conditional target factuality of 90%, and in Figure 16b, CF is evaluated on Wikipedia Biographies for a conditional target factuality of 90%. In Figure 16c, CF and pCF with frequency scoring are evaluated on MATH for marginal target factualities of 90% and 99%, respectively, and in Figure 16d, CF and pCF with frequency scoring are evaluated on FActScore and Natural Questions for a marginal target factuality of 90%. On MedLFQA for marginal factuality, CF meets the target, but for conditional factuality on both MedLFQA and Wikipedia Biographies, it fails

by ≈2-3%. For claim retention on MedLFQA, CF retains ≈26% of claims, and on Wikipedia Biographies, ≈79% of claims; this indicates that performance can be very dependant on the dataset evaluated on. On MATH at both target factualities, FActScore, and Natural Questions, both CF and pCF achieved the target factuality in all instances. For claim retention, on MATH at 90% pCF retained 12% more claims than CF, but on MATH at 99% and Natural Questions, both methods retained the same number of claims. On FActScore, pCF actually retained 18% less claims than CF, indicating that whilst in some cases the partial factuality setting can improve claim retention, in others it either has no effect, or can degrade claim retention.

A limitation many conformal factuality methods is that they treat atomic sub-claims as if they can be judged independently. This is unreasonable for reasoning tasks such as mathematics, where a step may be locally true or false only relative to the claims that precede it. A conformal filter that erases an earlier step can leave later steps unsupported, such that the resulting answer may still satisfy 'independent factuality' while becoming logically incoherent. Another notion of factuality called 'coherent factuality' is proposed to preserve not only the truth of retained claims but also the internal structure of reasoning [67]. Consider the steps to solve an algebraic expression; independent factuality removes explicitly false steps but leaves a later algebraic step that references a deleted premise, producing a final answer that is no longer understandable or justified. Therefore, correctness in reasoning tasks should be judged at the level of an ordered argument, not as a bag of individually factual claims.

In the experiments outlined in Section 5, coherent factuality with sub-graph filtering using graph-independent scoring (denoted coherent-SGF) and coherent factuality using descendants weight boosting (denoted coherent-DWB) are evaluated in Figure 16c. On MATH at a target factuality of 90%, both coherent variants meet the target, but coherent-DWB retains 11% less claims than coherent-SGF. However, on MATH at a target factuality of 99%, coherent-DWB retains 29% more claims than coherent-SGF, whilst both coherent variants meet the target factuality. This is significant because conformal guarantees can always be achieved trivially by rejecting almost all claims; this prove that graph-guided filtering avoids this degenerate regime. On MATH we compare the coherent variants to conformal factuality (denoted CF) and partial conformal factuality (denoted pCF) both with frequency scoring.

Coherent factuality has some limitations. Firstly, the guarantee is marginal validity rather than conditional, and it assumes exchangeability and therefore is not robust to distribution shift. Secondly, the inherent subjectivity of ground-truth and deduction is limiting; whether a step counts as justified depends on the annotator and the problem context. A simple arithmetic manipulation might be implicit in one setting and essential in another, and similarly, what can be assumed from the question context varies across domains. Therefore, 'coherent factuality' is not an absolute property of a reasoning trace but one defined relative to a human judgement protocol. While philosophically unavoidable, it does constrain the universality of the guarantee.

(b) Group-conditional Conformal Factuality

Conditional boosting and level-adaptive conformal prediction [79] are proposed to address two of these limitations of conformal factuality [65]. Firstly, the guarantees are only marginal, and at high factuality often remove many valid claims due to imperfect scoring functions. Secondly, it provides formally valid but overly censored outputs that have limited user value. The proposed methods extend conformal factuality to group-conditional validity and introduce score optimisation for better claim retention.

The authors model an LLM response R_i for a prompt P_i as a set of parsed claims $C_i =: \{C_{ij}\}_{j=1}^{k_i}$ with binary labels W_i indicating the underlying factuality of each claim. A filtered output is then defined by retaining only claims whose confidence scores exceed a calibrated threshold. Specifically, the filtering rule is $\hat{F}(C_i) := F(C_i; \hat{\tau}) = \{\, C_{ij} : p(P_i, C_{ij}) \geq \hat{\tau} \,\}$, where p is a claim-scoring function and $\hat{\tau}$ is the $\frac{\lceil (1-\alpha)(n+1) \rceil}{n+1}$-quantile of the conformity scores on a calibration set of n prompt-response-claim-annotation tuples, $\{(P_i, R_i, C_i, W_i)\}_{i=1}^{n}$. For a monotone loss $\mathcal{L}$, the conformity score is $s(C_i, W_i) = \inf\{\, \hat{\tau} : \mathcal{L}(F(C_i; \hat{\tau}), W_i) \leq \lambda \}$ for a user-chosen tolerance λ.

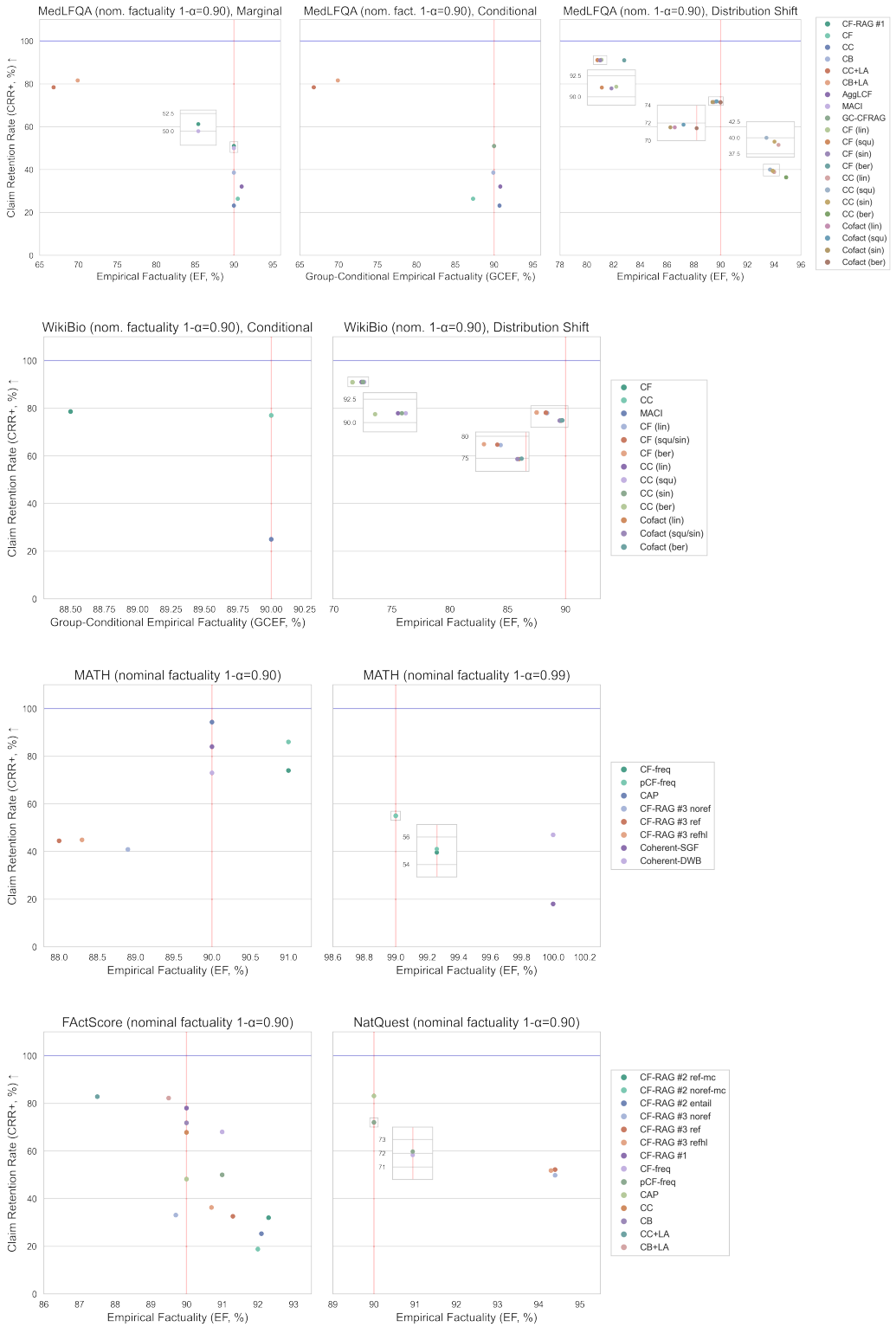

Figure 16: Scatterplots for conformal factuality methods on MedLFQA, Wikipedia Biographies, MATH, FActScore, and Natural Questions. Both marginal and conditional factuality at various targets are explored, including scenarios under distribution shift. Metrics are defined in Section 5. The x-axis is empirical factuality, and the red line indicates the target factuality; to the left of this line is undercoverage and to the right is overcoverage. The y-axis is the % claims retained; higher values are better, and the blue line indicates maximum claim retention for a perfect method.

Exact distribution-free covariate-conditional validity is impossible without vacuous outputs, so the function-class-based notion, quasi-conditional conformal prediction (QCCP) [38], is adopted. Rather than requiring validity conditional on every covariate, the guarantee is enforced over a chosen linear class $\mathcal{F} = \{\Phi(X)^\top \beta : \beta \in \mathbb{R}^d\}$, where $X = \Phi(P, R)$ are prompt/response-derived features. Therefore, to obtain conditional guarantees, the augmented quantile-regression is fit over $\mathcal{F}$ with pinball loss, yielding a cut-off $\hat{\tau}(x_{n+1})$ that satisfies

$$\mathbb{E}\Big[f(x_{n+1})\left(\mathbf{1}(\mathcal{L}(\hat{F}(C_{n+1}), W_{n+1}) \le \lambda) - (1-\alpha)\right)\Big] = 0, \quad \forall f \in \mathcal{F}, \tag{10.1}$$

which yields group-conditional validity at level α if $\mathcal{F}$ is the span of group indicators. For the level-adaptive procedure, this is generalised to a prompt-dependent level $\alpha(\mathcal{X})$, obtaining

$$\mathbb{E}\Big[f(x_{n+1})\left(\mathbf{1}(\mathcal{L}(\hat{F}(C_{n+1}), W_{n+1}) \le \lambda) - (1-\alpha(x_{n+1}))\right)\Big] = 0. \tag{10.2}$$

Rather than fixing $\alpha(\cdot)$, this allows the claimed correctness probability to vary across prompts while remaining empirically calibrated. However, $\alpha(\cdot)$ is not known ahead of time and so must be estimated on a separate fold of the data, learning the minimum error level needed to satisfy a target utility constraint, such as minimum claim retention. The level-adaptive procedure is then run on a second data fold using the learned α. Conditional boosting is introduced to improve claim retention. It is a differentiable optimisation of the scoring function p_θ (a parameterised claim-scoring function that assigns a measure of confidence to each sub-claim) through the conformal quantile-regression program. The parameter θ is optimised to maximise retained claims on the holdout data set of size m after conditional conformal calibration, $\theta^* = \arg\max_\theta \sum_{i=1}^{m} \sum_{j=1}^{k_{n+i}} \mathbf{1}(p_\theta(P_{n+i}, C_{(n+i)j}) \ge \hat{\tau}_i)$, where $\hat{\tau}_i$ is the filtering threshold output by the conditional conformal method on the held-out test point x_{n+i}. For a linear function $\mathcal{F}$, the threshold $\hat{\tau}_i(\theta)$ is shown to be differentiable through the optimal linear program basis. The authors run a training loop in the spirit of ConfTr [213], repeatedly splitting data into temporary calibration/test subsets, calibrating conditionally, evaluating retention, and updating θ by gradient descent. Empirically, boosting is shown to improve claim retention.

In the experiments outlined in Section 5, conditional conformal, conformal boosting, conditional conformal with level-adaptive, and conditional boosting with level-adaptive are denoted CC, CB, CC+LA, and CB+LA, respectively. In Figure 16a, all methods are evaluated on MedLFQA for both a marginal and conditional target factuality of 90%, and in Figure 16d, all methods are evaluated on FActScore with a marginal target factuality of 90%. In addition, CC is evaluated in Figure 16b on Wikipedia Biographies for a conditional target factuality of 90%. On MedLFQA for both marginal and conditional factuality, CC and CB meet the target, with CB retaining $\approx 15\%$ more claims. The level-adaptive variants retained many more claims, with CB+LA retaining 43% more than CB and $\approx 9\%$ more than CC. However, these gains in claim retention came at the cost of lowering the factuality to $\approx 70\%$ on average, 20% below target. On FActScore, both CC and CB meet the target factuality, but the gains of CB over CC were marginal at +4% claims retained. For CB+LA, this increased to around 10% claims retained, but unlike with MedLFQA, factuality is close to the target this time.

Both proposed methods have their limitations; firstly, they are not robust to distribution shift due to the exchangeability assumption. Secondly, the conditional validity is only enforced relative to the chosen function class F, not universally; this means that a weak feature class can lead to spuriously calibrated nominal levels, or a rich feature class may improve calibration at the cost of computation and efficiency. Lastly, the optimisation loop for conditional boosting is computationally heavy, as it repeatedly reruns conditional conformal calibration during training.

(c) Distribution-shift-aware Conformal Factuality

So far, prior conformal factuality methods for LLMs share a common fundamental weakness; their guarantees depend on exchangeability between calibration and test data [65,67,79,130], but in practice prompts drift over time; prompts change in both style and difficulty, user

populations change, topic distributions change, and prompts arrive sequentially from a non-stationary process. Under such a covariate shift, nominal hallucination control can be lost by split-conformal-based factuality methods. Therefore, there is a need for a method which can preserve factuality guarantees for LLM-generated outputs when prompts arrive online from an unknown and continually evolving distribution, when no ground-truth claim labels are revealed after each prediction. CoFact is proposed to meet this need [137]. Let $z_i = (P_i, R_i, C_i)$, $z_i \in \mathcal{Z}$ represent the prompt-response pair for sample i, where P_i is the prompt, R_i is the corresponding response generated by the LLM, and $C_i = \{C_{ij}\}_{j=1}^{k_i}$ denotes the set of claims extracted from R_i. Each sample in the calibration set can be represented as (z_i, W_i) where $W_i = \{W_{ij}\}_{j=1}^{k_i}$, $W_i \in \{0, 1\}$ represents the binary factuality labels for each claim C_{ij} with $W_{ij} = 1$ if C_{ij} is factual, otherwise $W_{ij} = 0$. CoFact assumes pure covariate shift, meaning that the conditional distribution remains invariant $\mathcal{D}_t(\mathcal{W} \mid \mathcal{Z}) = \mathcal{D}_0(\mathcal{W} \mid \mathcal{Z})$ where $\mathcal{D}_0$ and $\mathcal{D}_t$ represent calibration distribution and test distribution respectively, with bounded density ratio $r_t^*(\mathcal{Z}) = \mathcal{D}_t(\mathcal{Z})/\mathcal{D}_0(\mathcal{Z})$, rather than exchangeability between calibration and test data. In the oracle setting, the conformal threshold is computed using importance-weighted calibration scores w_t given the conformity scores obtained from the calibration set $\{z_i\}_{i=1}^n$ and the test sample z_{n+t} at any round $t \in \{1, ..., T\}$

$$\begin{aligned} \tau_t &= \text{Quantile}\left(\sum_{i=1}^{n} w_t^i(z_i)\,\delta_{z_i} + w_t^i(z_{n+t})\,\delta_\infty;\ 1-\alpha\right), \\ w_t^*(z) &= \frac{r_t^*(z)}{\sum_{i=1}^n r_t^*(z_i) + r_t^*(z_{n+t})}, \end{aligned} \tag{10.3}$$

where δ_s denotes the Dirac measure at $s \in \mathbb{R}$. Filtered responses are then given by $\hat{F}_t(C_{n+t}) = \{C_{(n+t)j} \in C_{n+t} : p(C_{(n+t)j}, P_{n+t}) \geq \hat{\tau}_t\}$.

The core practical problem is therefore to estimate the true density ratio r_t^* online, as it is unknown at time t and test points arrive sequentially and without labels. CoFact adopts the online density-ratio estimation (ODRE) framework [218] formulated as dynamic-regret minimisation over a logistic-ratio model $r_t(z; \theta_t) = \exp(-\phi(z)^\top \theta_t)$, with Bregman-divergence objective, rather than over conformal thresholds directly. The empirical optimisation target becomes $\min_\theta \sum_{t=1}^T \hat{\mathcal{L}}_t(\theta) - \hat{\mathcal{L}}_t(\theta_t^*)$, with $\hat{\mathcal{L}}_t(\theta) = \mathbb{E}_{z\sim D_0}\left[\partial\psi(r(z;\theta))\, r(z;\theta) - \psi(r(z;\theta))\right]$ $- \mathbb{E}_{z\sim D_t}\left[\partial\psi(r(z;\theta))\right]$, where θ_t^* is the optimal model parameter corresponding to the true density ratio r_t^*, and $\psi(r)$ is the divergence function $\psi(r) = r\log r - (r+1)\log(r+1)$.

In the experiments outlined in Section 5, conformal factuality (denoted CF), conditional conformal (denoted CC), and CoFact are evaluated in Figures 16a and 16b, with covariate shift simulated by mixing two base distributions according to four temporal patterns: linear drift (lin), square-wave periodic switching (squ), sine-wave periodic drift (sin), and Bernoulli switching (ber). On MedLFQA for a marginal target factuality of 90%, CF failed significantly to meet the target, by ≈ 7-9% with a high avg. claim retention of ≈ 91%. CC, on the other hand, overshot the target by ≈ 4-5%, but had a low claim retention avg. of ≈ 40%. In comparison, CoFact achieved the target factuality with an avg. claim retention of ≈ 72%. Whilst fewer overall claims were retained than CF, more of these claims are likely to be correct due to meeting the target factuality. On Wikipedia Biographies for a marginal target factuality of 90%, CC underperformed significantly; whilst it had an avg. claim retention of 91%, it failed to meet the target by ≈ 18%. CF was closer to the performance of CoFact on this dataset, with CF only failing to meet the target by ≈ 2-3% with a 78% avg. claim retention, compared to CoFact meeting the target with a 75% avg. claim retention. Despite this, CoFact still remains the only method to consistently meet the target factuality, and appears to be highly stable under smooth, periodic, and abrupt variants of covariate shift.

CoFact is not without its limitations, however. The principal technical limitation is the assumption of pure covariate shift.This excludes settings where the relationship between prompt-response content and factuality changes over time, for example, if the claim-scoring model degrades on newly emerging topics or if the ground-truth knowledge source itself evolves, known as concept drift. Another limitation is that the guarantee is not the standard per-example conformal coverage guarantee, rather it controls the deviation of the average hallucination rate

over time from α. This is acceptable for online unlabelled streams, but does not ensure that each individual answer or local time segment satisfies the target error rate. This is demonstrated on their WildChat+ benchmark, where the first 50 steps explicitly show a burn-in phase where adaptation is not yet efficient. CoFact is a workable approximation, but not fully distribution-free.

(d) Conformal Factuality with Retrieval Augmented Generation

Prior conformal factuality methods for LLMs operate largely from the model's parametric knowledge and do not exploit the extra structure available in retrieval-augmented generation (RAG) [65,67,79,130,137]. As such, conformal factuality with retrieval augmented generation (CF-RAG) is proposed to adapt conformal factuality to the RAG setting by using retrieved-document information to score and filter response sub-claims more intelligently, thereby keeping more useful content while preserving factuality guarantees [111]. CF-RAG also extends this to a group-conditional setting for multiple sub-domains.

Given a query $x \in \mathcal{X}$, the system retrieves a set of m documents $D = \{d_1, \ldots, d_m\}$, and then generates an answer $\hat{y}$ composed of p sub-claims, $\hat{y} = \{c_1, \ldots, c_p\}$. The goal is to post-process $\hat{y}$ by filtering out low-quality sub-claims to produce a refined answer $y^* \subseteq \hat{y}$ that is entailed by the ground truth answer y with probability at least $1 - \alpha$. An LLM is first used to decompose a generated answer into atomic sub-claims. Then, for sub-claim $c \in \hat{y}$, the method computes the sub-claim relevance score $R(c_k)$ for a query x. The scoring function $R(c)$ is the maximum over retrieved documents of the product between query-document similarity and claim-document similarity, truncated below at zero. The intuition is that a claim should be considered more reliable if it is semantically aligned with a retrieved document that is itself strongly relevant to the query. Next, calibration uses an automatic annotation function $A(c, x, y, D) \in \{0, 1\}$, implemented with an LLM given the query x, ground truth answer y, retrieved documents D, and generated sub-claim c. The annotation function $A(c, x, y, D) = 1$ when the sub-claim c is factual and $A(c, x, y, D) = 0$ otherwise.

For marginal conformal factuality, the method defines the nested filtering function $F_q(\hat{y}) = \{\, c \in \hat{y} : R(c) \geq q \,\}$. The threshold q is estimated using CP calibration with the conformal scores defined by $s(x_i, y_i) = \inf \left\{ q \in \mathbb{R}_+ : \forall q' \geq q,\ \forall c \in F_{q'}(\hat{y}_i),\ A(c, x_i, y_i, D) = 1 \right\}$. The score s is the smallest threshold $\hat{q}$ such that all retained claims are considered factual by the annotation function A. The estimated threshold $\hat{q}$ is the split-conformal $\lceil (n+1)(1-\alpha) \rceil / n$-quartile of calibration conformal scores for a calibration set of n examples. At test time, the returned answer is $y^*_{\text{test}} = F_{\hat{q}}(\hat{y})$, yielding the factuality target $\mathbb{P}(y^*_{\text{test}} \Rightarrow y_{\text{test}}) \geq 1 - \alpha$.

For the conditional extension, CF-RAG adopts Mondrian conformal prediction (see Section 4); calibration data is partitioned by group $g(x)$, and a separate threshold $\hat{q}_a$ is calibrated for each group a, then the threshold for $a_{\text{test}} = g(x_{\text{test}})$ is applied for the test example's group to get $y^*_{\text{test}} = F_{\hat{q}_{\alpha_{\text{test}}}}(\hat{y})$. As each group is calibrated independently, it is argued that the groupwise version of the marginal guarantee implies the group-conditional factuality. This differs from the learned pointwise adaptive thresholds of conditional conformal factuality [79], explored in Section 10(b).

In the experiments outlined in Section 5, CF-RAG and group-conditional CF-RAG are denoted CF-RAG #1 and GC-CFRAG, respectively. In Figure 16a, both methods are evaluated on MedLFQA for both a marginal and conditional target factuality of 90%, and in Figure 16d, CF-RAG is evaluated on FActScore with a marginal target factuality of 90%. On MedLFQA, CF-RAG/GC-CFRAG meets the target whilst maintaining the highest claim retention rate out of non-level-adaptive methods. This is expected, as the generated response contains claims which are strengthened by the evidence provided by the RAG mechanism, indicating that a RAG extension for conformal factuality is beneficial. On FActScore, CF-RAG meets the target whilst maintaining the highest claim retention rate out of non-level-adaptive methods, as expected. CF-RAG outperforms the three other RAG conformal factuality methods by $\approx 50\%$ on average.

CF-RAG has some limitations. For example, the relevance score is only a proxy for factuality; a claim can be factually correct but weakly supported by the retrieved documents, or strongly

semantically aligned with a retrieved document that is itself incomplete or misleading. Lastly, CF-RAG relies on standard exchangeability assumptions.

Table 5: Performance of conformal factuality methods. EF, GCCF, and CRR+ are reported as percentages; for EF and GCCF closer to the target factuality is better, and CRR+ $= 100\%$ is best; for definitions see Section 5, Table 2. Where possible, the error rate used is $\alpha = 0.10$ (nominal factuality $1 - \alpha = 0.90$), but this may not be for all cases; please refer to the performance notes.

Ref	Conformal Method	EF	CRR+↑	GCCF	Performance Notes
[65]	Conformal factuality	89.3	62.0	87.3	Mean across 3 datasets using frequency scoring. GCEF: mean of boxplot medians across MedLFQA (5 datasets) from [130].
[79]	Conditional Conformal	90.0	23.2	90.7	Mean of boxplot medians across MedLFQA (5 datasets). EF/GCEF: from [130].
[79]	Conditional Boosting	90.0	38.6	89.9	Same as above.
[79]	CC+LA	66.8	78.4		EF: unweighted mean realised level across nominal bins (from Figure 9); CRR+: at adaptive nominal $1 - \alpha(x) \in \{0.50, 0.85\}$.
[79]	CB+LA	69.9	81.6		Same as above.
[67]	Coherent factuality	91.0	85.8		Mean across 2 datasets using sub-graph filtering.
[110]	MVSC	90.4	10.8	[86.2, 93.8]	Mean across 2 LLMs $\times$ 2 datasets using self-consistency scoring. CRR+ = mean retained facts per biography / mean original claims per biography. Original claims per biography sourced from Table 15. GCEF $\in [0.90 - e, 0.90 + e]$, where e is mean absolute groupwise coverage error.
[110]	GCCQR	90.4	9.4	[86.1, 93.9]	Same as above.
[111]	Conformal-RAG	90.5	65.5	90.0	Mean across 4 datasets. GCEF: mean across 2 datasets (Wiki = PopQA + HotpotQA).
[112]	Conformal-RAG	92.3	33.0		Using reference model confidence. CRR+ $=$ (TPR $\times$ FPR)/(TPR $\times$ $(1 -$ EF$)$ + EF $\times$ FPR) $\times$ 100.
[143]	Conformal-RAG	91.3	41.3		Mean across 4 LLMs $\times$ 3 datasets with highlight using numeric score. CRR+ = same formula as above.
[130]	AggLCF	91.0	32.1	90.8	Mean of boxplot medians across MedLFQA (5 datasets). CRR+=(mean retained sub-claims / cluster count) $\times$ 100. Cluster count = 10.07 after multi-model aggregation and clustering.
[157]	ConfLVLM	89.3	4.3		Mean across 3 LVLMs on MIMIC-CXR using BiomedCLIP scoring with tolerance $\lambda = 0$ (strict setting).
[137]	CoFact-lin	89.6	73.2		Mean across 2 datasets for linear shift.
[137]	CoFact-squ	89.6	73.3		Same above for square shift.
[137]	CoFact-sin	89.5	73.2		Same above for sine shift.
[137]	CoFact-ber	89.9	73.2		Same above for Bernoulli shift.
[137]	CoFact-wild	92.9	14.7		Mean across all time steps on WildChat+.
[148]	MACI	90.0	30.0	90.3	Mean across 3 datasets. GCEF: mean across 3 groups within 3 datasets.
[89]	CAP	90.0	75.2		EF not tabulated but maintains desired factuality. CRR+: mean across 3 datasets.

11. Conformal Retrieval Augmented Generation

This section covers conformal retrieval-augmented generation (RAG) methods illustrated in Figure 2. Conformal RAG methods are also applied to the factuality setting, explored in Section 10(d). Figure 17 is an example where conformal RAG is effective in reducing hallucinations.

Retrieval-augmented generation (RAG) systems are often presented as a remedy for LLM hallucinations, as they retrieve supporting evidence from a larger knowledge base before generation. However, in standard RAG systems, there remains no formal guarantee that the final answer generated by the LLM, even when informed by the returned contents of the RAG system, will actually be correct. Two distinct failure modes are identified; the retriever may fail to return a relevant document or passage, and the generator (LLM) may still produce an incorrect answer even when given relevant evidence. Existing conformal methods either focused on close-ended settings, required access to token probabilities, or handled only the generator (LLM) and not the retrieval component. Thus, trustworthy retrieval augmented question answering (TRAQ) is

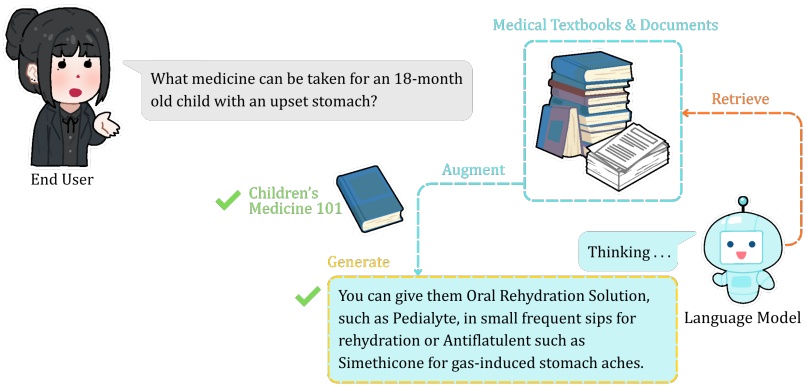

Figure 17: **Conformal Retrieval Augmented Generation** addresses tasks for which the language model's parametric knowledge is insufficient. In such cases, relevant evidence from an external knowledge source is used to generate a grounded response rather than a hallucinated one. The green check mark indicates that the answer is correct and that the correct evidence was retrieved.

proposed to provide an end-to-end statistical guarantee for open-ended RAG, where correctness means the final answer set contains a semantically correct answer with high probability [60]. A second motivation is that open-ended question-answering (OpenQA) has many semantically equivalent ways to express the same answer, which causes problems for standard probability-based confidence measures, as confidence can be fragmented across paraphrases. Therefore, TRAQ should operate at the semantic level rather than the level of exact strings. TRAQ is a conformal method designed for retrieval-augmented generation (RAG) systems. It treats the two main parts of a RAG pipeline separately: the retriever, which selects relevant documents, and the generator, which produces answers using those documents.

First, for the retrieval stage, TRAQ evaluates how well each passage matches the query using a similarity score between the question and each candidate passage. It then applies conformal calibration to decide which passages are "reliably relevant". The result is a set of candidate passages that is guaranteed (up to a chosen error level) to contain the truly relevant passage with high probability. Second, for the generation stage, TRAQ considers each selected passage separately. For each question–passage pair, the language model generates multiple possible answers by sampling. These answers are then grouped into clusters based on semantic similarity, so that different phrasings of the same idea are treated as one response. Clusters that appear more frequently among the samples are considered more reliable. Conformal calibration is then applied again to keep only a set of answers that is statistically reliable for each passage. Finally, TRAQ combines the results from all retained passages. It takes all the answer sets produced for each passage and merges them into one final answer set. Because both retrieval and generation are individually controlled for error, the overall system retains a guaranteed reliability level determined by the sum of the two error budgets.

To improve efficiency, TRAQ uses Bayesian optimisation to reduce the average prediction set size (APSS) while preserving the theoretical guarantee. It uses a held-out optimisation set and a Gaussian-process-based Bayesian optimisation loop to search for the best error-budget allocation. This is achieved by repeatedly sampling candidates subject to error budgets assigned to the retriever and the language model generator, normalising them so they sum to the target overall error level α, constructing prediction sets on the optimisation split, evaluating a performance metric, updating the Gaussian process, and retaining the best allocation. A variant of TRAQ, which provides a probably approximately correct (PAC) [46] guarantee, is called TRAQ-P.

In the experiments outlined in Section 5, TRAQ and TRAQ-P are evaluated in Figure 6b. On TriviaQA at a target coverage of 90%, both TRAQ and TRAQ-P have significant overcoverage of $\approx$ 8-9%, with high APSS, though TRAQ has a lower APSS than TRAQ-P. On TriviaQA at a target coverage of 70%, both TRAQ and TRAQ-P have more significant overcoverage of $\approx$ 15-20%, but much lower APSS than before, though TRAQ still has a lower APSS than TRAQ-P. The significant overcoverage observed suggests that these methods are very conservative. TRAQ has some limitations. Firstly, the TRAQ depends on two strong task-specific assumptions; retrieval must place a relevant passage in the top-K, and generation must include a semantically correct answer in the top-M samples. As the conformal guarantee only applies to these candidates, failures in either stage can undermine its practical usefulness; increasing K or M may mitigate this, but at added cost. Secondly, TRAQ depends on semantic clustering quality. If the Rouge-based or NLI-based clustering groups unrelated answers together, or fails to merge equivalent ones, both the uncertainty measure and the final set size can be distorted. Thirdly, TRAQ has substantial computational overhead, as it retrieves multiple passages and then samples multiple responses per passage; the retrieval-phase cost grows approximately linearly with the number of retrievals.

A practical weakness of RAG is that retrieved context is often too long and too noisy, yet most pre-generation filtering methods rely on heuristic cut-offs or uncalibrated similarity scores. Motivated by the need for a principled context-engineering method that can reduce prompt clutter while guaranteeing that supporting evidence is not discarded at an uncontrolled rate, a split-conformal-based framework is proposed for snippet filtering after retrieval [165]. Retrieved documents are segmented into overlapping snippets, each assigned a nonconformity score using either an embedding-based scorer or an LLM-based relevance scorer, where lower scores indicate greater relevance. Experiments show that conformal filtering consistently meets its target coverage while shrinking retained context by roughly 2-3$\times$. A limitation is that the guarantee is only as good as the relevance labels used in calibration, and provides only marginal coverage and depends on the exchangeability assumption, which may be violated under distribution shift.

12. Conformal Risk Control Methods

This section covers alternative conformal risk control (CRC) methods illustrated in Figure 2.

A central problem in multi-LLM deployment is that although SLMs are cheaper than LLMs, existing routers often underused them because their capability boundaries are difficult to predict. This manifests in two linked issues; first, standard embeddings capture semantic similarity but not whether a specific SLM can actually answer a query; second, most routing methods rely on heuristic thresholds and therefore lack a principled way to control routing risk while balancing accuracy against cost. Conformal risk-controlled routing (CR^2) is proposed to address both of these issues [142]. CR^2 (illustrated in Figure 18) decomposes routing into two stages; in stage one, the router decides whether the smallest SLM can handle the query, using a capability-aware representation learned through supervised contrastive learning (SCL). This reshapes the embedding space so that queries the SLM can answer correctly are separated from those it cannot, even when they are semantically similar. In stage two, deferred queries are scored by a multi-label classifier over the LLM pool, producing a candidate set of plausible LLMs. A global threshold for this candidate set is then calibrated using CRC on a bounded, monotone composite loss that combines stage one errors with the candidate-set model-level false-positive rate in stage two. The final router chooses the lowest-cost model within the calibrated candidate set.

Another weakness in multi-LLM deployment is that most existing routers make a single-model selection, so an incorrect routing decision can substantially degrade downstream performance. As such, there is a need for a routing strategy that is both efficient and statistically reliable; one that can select a small subset of candidate LLMs, or abstain entirely, while guaranteeing that at least one suitable LLM is included with high probability. Risk-aware calibrated efficient routing (RACER) is a routing framework proposed to handle this weakness [163]. RACER reformulates routing as the α-valid optimal routing (α-VOR) problem; minimise the expected size of the routed LLM set subject to a user-specified upper bound α on misrouting risk, defined as the event

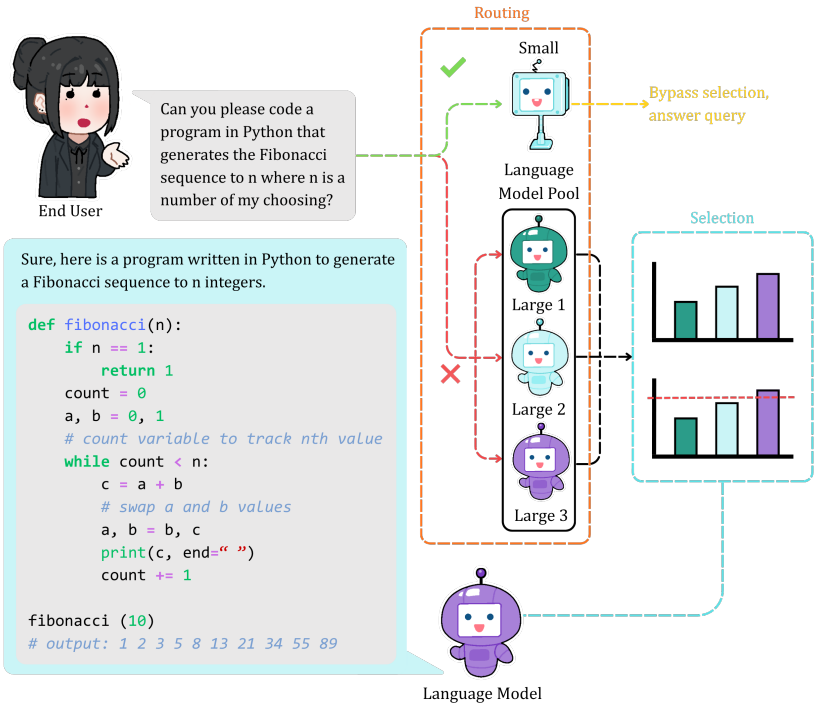

Figure 18: **Routing LLMs** can be useful when one is looking to optimise costs whilst maintaining high accuracy. In the example here based on [142], the router takes a prompt and evaluates whether an SLM has the capability to adequately answer a question; if not, deferred queries are scored by over an LLM pool, producing a candidate set of plausible LLMs from which a global threshold is calibrated, selecting the lowest-cost model within the set to answer the query.

that the selected set excludes all correct LLMs. RACER augments the LLM pool with a virtual null model to represent abstention, constructs nested LLM sets using augmented router scores and a nonconformity function, and then calibrates a threshold with a finite-sample conformal procedure. At inference time, the resulting calibrated set can be aggregated through majority voting or weighted aggregation, allowing complementary strengths of multiple LLMs to be exploited without evaluating the full ensemble.

Another common deployment problem for LLMs is the balancing of competing objectives, such as cost vs. accuracy or helpfulness vs. harmlessness; existing approaches require retraining and only fix the LLM at a single trade-off point, which is undesirable. Motivated by the need for a post-hoc, black-box method that can tune this trade-off after deployment while providing explicit statistical guarantees that a chosen guardrail metric remains within budget, conformal arbitrage is proposed [131]. Conformal arbitrage uses two pre-existing systems; a Primary model optimised for the main objective, and a more conservative Guardian aligned with the guardrail objective. A score-gap threshold is constructed: if the Primary model's top option is sufficiently far ahead of its alternatives, the system accepts it; else, it defers to the Guardian. This threshold is calibrated using conformal risk control, yielding a finite-sample, distribution-free guarantee that the expected guardrail loss does not exceed a user-specified level α.

13. Conclusion

This paper presents the first review of conformal prediction for large language models (LLMs), with in-depth and extensive analysis of over 106 papers, covering the full state-of-the-art from the inception of this field to the start of 2026, with a companion repository available online [20]. We proposed a novel taxonomy with six distinct categories to represent the current landscape of research and practice, using this data to observe trends and patterns. Based on the trends analysed in Sections 6(a) and 6(b), we provided two recommendations to govern research going forward.

First, we recommend that more black-box LLMs be used in experiments, as we found that out of 224 unique LLMs, only 19% were black-box models. We make a case that improving the reliability of black-box LLMs, particularly frontier LLMs such as ChatGPT and Gemini that have millions of global active users, will have **the greatest impact on human society**. At the same time, further work on conformal prediction for black-box LLMs is needed. Although several logit-free methods have been proposed (see Section 7(b)), important challenges remain only partially resolved. In open-ended generation, finite-sample approximation of the output space can be expensive, and performance often improves with additional sampling budget. Existing methods rely on heuristic uncertainty scores and semantic similarity models that may be noisy or misspecified, so calibration quality is limited by the quality of these external signals. Some methods require clustering procedures which add additional computational overhead, and guarantees remain marginal and depend on exchangeability which can be violated under distribution shift. We believe that through developing methods for black-box models, **better notions of uncertainty than miscalibrated logits and log-probabilities may be found**, which could benefit the entire LLM landscape. We think that broader empirical validation of state-of-the-art proprietary models is therefore an important direction for future work.

Second, we recommend that more task-wise variation be adopted in the validation of conformal methods with LLMs, through **a more diverse selection of datasets including lesser known benchmarks**, rather than researchers optimising their methods on a single popular benchmark which may introduce bias. We also **recommend that a minimum of four datasets be used** in evaluation, as many conformal studies are reported with less, sometimes with just one dataset [50,76,112,115,130,132,146], and it is not known how well the model generalises to other tasks, domains, and data distributions. We give examples of how typical evaluation pipelines can be enhanced with more task variety, i.e. swapping out three similar open-domain QA datasets for open-domain QA datasets which also validate reading comprehension, multi-step reasoning, and commonsense inference. We also propose that, in some cases where benchmarks are specialised and the options to alternate are narrower, just supplementing a pipeline with **an additional dataset is enough to improve task diversity and model generalisability**.

We also conducted subgroup studies using a diverse subset of methods evaluated in a controlled performance analysis.In practice, we observed logit-free methods outperforming logit-based methods, with **lower mean absolute coverage error and prediction set size**. This indicates that our hypothesis did not hold, and in fact, we observed the opposite to be true. We consider two primary drivers for this finding; that those black-box uncertainty techniques **may be better-aligned conformity signals**, and that they **may sidestep known issues with miscalibrated token probabilities**. We also observed conformal methods for LVLMs **exhibit higher overcoverage error and larger prediction set sizes** as expected, which confirms our hypothesis, though we cannot fully attribute this to additional sources of uncertainty in visual-language tasks. Therefore, while we do not claim a definitive causal explanation, the empirical evidence indicates that, in our experiments, **conformal methods for LVLMs exhibit a stronger coverage-informativeness trade-off than those for LLMs**.

Data Accessibility. A companion repository for this review is available at [20].

Funding. This work is supported by EPSRC through the Centre for Doctoral Training in AI for Digital Media Inclusion (Grant No. EP/Y030915/1). Khuong An Nguyen is supported by UKRI DAFNI Fellowship grant.

+

References

1. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. 2023 Large language models in medicine. *Nature Medicine* **29**, 1930–1940.
2. Zeng F, Gan W, Wang Y, Liu N, Yu PS. 2023 Large language models for robotics: A survey. *arXiv preprint arXiv:2311.07226*.
3. Lai J, Gan W, Wu J, Qi Z, Yu PS. 2024 Large language models in law: A survey. *AI Open* **5**, 181–196.
4. Chhikara P. 2025 Mind the Confidence Gap: Overconfidence, Calibration, and Distractor Effects in Large Language Models. *Transactions on Machine Learning Research (TMLR)*. https://openreview.net/forum?id=lyaHnHDdZl.
5. Huang L, Yu W, Ma W, Zhong W, Feng Z, Wang H, Chen Q, Peng W, Feng X, Qin B et al.. 2025 A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems* **43**, 1–55.
6. Gallegos IO, Rossi RA, Barrow J, Tanjim MM, Kim S, Dernoncourt F, Yu T, Zhang R, Ahmed NK. 2024 Bias and fairness in large language models: A survey. *Computational Linguistics* **50**, 1097–1179.
7. Das BC, Amini MH, Wu Y. 2025 Security and privacy challenges of large language models: A survey. *ACM Computing Surveys* **57**, 1–39.
8. Weidinger L, Mellor J, Rauh M, Griffin C, Uesato J, Huang PS, Cheng M, Glaese M, Balle B, Kasirzadeh A et al.. 2021 Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.
9. Vovk V, Gammerman A, Shafer G. 2005 *Algorithmic Learning in a Random World*. Springer.
10. Shafer G, Vovk V. 2008 A tutorial on Conformal Prediction.. *Journal of Machine Learning Research (JMLR)* **9**.
11. Papadopoulos H, Proedrou K, Vovk V, Gammerman A. 2002 Inductive confidence machines for regression. In *Proceedings of the 13th European Conference on Machine Learning (ECML)* pp. 345–356. Springer.
12. Angelopoulos AN, Bates S. 2023 Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning* **16**, 494–591.
13. Nguyen KA, Luo Z. 2018 Cover your cough: Detection of respiratory events with confidence using a smartwatch. In *7th Workshop on Conformal and Probabilistic Prediction and Applications (COPA)* vol. 91 pp. 114–131. Proceedings of Machine Learning Research (PMLR).
14. Obayemi A, Nguyen KA. 2025 Uncertainty quantification of multimodal models. In *Proceedings of the 38th International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems (IEA/AIE)* pp. 272–280. Springer.
15. Ashby AE, Meister JA, Nguyen KA, Luo Z, Gentzke W. 2022 Cough-based COVID-19 detection with audio quality clustering and confidence measure based learning. In *11th Symposium on Conformal and Probabilistic Prediction with Applications (COPA)* vol. 179 pp. 129–148. Proceedings of Machine Learning Research (PMLR).
16. Meister JA, Nguyen KA. 2025 Conformalised data synthesis. *Machine Learning* **114**, 57.
17. Choudhury R, Luo Z, Nguyen KA. 2023 Evaluating potential sensitive information leaks on a smartphone using the magnetometer and Conformal Prediction. In *12th Symposium on Conformal and Probabilistic Prediction with Applications (COPA)* vol. 204 pp. 116–133. Proceedings of Machine Learning Research (PMLR).
18. Riquelme-Granada N, Nguyen KA, Kapetanakis S, Luo Z. 2025 Calibrated large language models for multi-label classifications. US Patent 12,367,223.
19. Kapetanakis S, Nguyen KA, Riquelme-Granada N, Luo Z. 2025 Reliable outputs from large language models for multi-label classification tasks. US Patent 12,499,134.
20. Ashby AE. 2026 Conformal LLM Review: Companion Repository for "Uncertainty-Aware Large Language Models: A Scoping Review of Conformal Prediction Methods". Available at: https://github.com/aliceinlatentspace/conformal-llm-review.
21. Moher D, Liberati A, Tetzlaff J, Altman DG, Group P et al.. 2010 Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *International Journal of Surgery* **8**, 336–341.
22. Campos M, Farinhas A, Zerva C, Figueiredo MA, Martins AF. 2024 Conformal prediction for natural language processing: A survey. *Transactions of the Association for Computational Linguistics (TACL)* **12**, 1497–1516.

23. Zhou X, Chen B, Gui Y, Cheng L. 2025 Conformal prediction: A data perspective. *ACM Computing Surveys* **58**, 1–37.
24. Shorinwa O, Mei Z, Lidard J, Ren AZ, Majumdar A. 2025 A survey on uncertainty quantification of large language models: Taxonomy, open research challenges, and future directions. *ACM Computing Surveys* **58**, 1–38.
25. Ji W, Yuan W, Getzen E, Cho K, Jordan MI, Mei S, Weston JE, Su WJ, Xu J, Zhang L. 2025 An overview of large language models for statisticians. *arXiv preprint arXiv:2502.17814*.
26. Guo C, Pleiss G, Sun Y, Weinberger KQ. 2017 On calibration of modern neural networks. In *34th International Conference on Machine Learning (ICML)* vol. 70 pp. 1321–1330. Proceedings of Machine learning Research (PMLR).
27. Liu H, Li C, Li Y, Lee YJ. 2024 Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* pp. 26296–26306.
28. Manokhin V. 2022 Awesome Conformal Prediction. (10.5281/zenodo.6467205)
29. Sadinle M, Lei J, Wasserman L. 2019 Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association (JASA)* **114**, 223–234.
30. Romano Y, Sesia M, Candès E. 2020 Classification with valid and adaptive coverage. *Advances in Neural Information Processing Systems* **33**, 3581–3591.
31. Angelopoulos AN, Bates S, Jordan M, Malik J. 2021 Uncertainty Sets for Image Classifiers using Conformal Prediction. In *9th International Conference on Learning Representations (ICLR)*. https://openreview.net/forum?id=eNdiU_DbM9.
32. Angelopoulos AN, Bates S, Fisch A, Lei L, Schuster T. 2024 Conformal Risk Control. In *12th International Conference on Learning Representations (ICLR)*. https://openreview.net/forum?id=33XGfHLtZg.
33. Snell J, Zollo TP, Deng Z, Pitassi T, Zemel R. 2023 Quantile Risk Control: A Flexible Framework for Bounding the Probability of High-Loss Predictions. In *11th International Conference on Learning Representations (ICLR)*. https://openreview.net/forum?id=p6jsTidUkPx.
34. Jin Y, Candès EJ. 2023 Selection by prediction with conformal p-values. *Journal of Machine Learning Research (JMLR)* **24**, 1–41.
35. Jin Y, Candès EJ. 2025 Model-free selective inference under covariate shift via weighted conformal p-values. *Biometrika*.
36. Angelopoulos AN, Bates S, Candès EJ, Jordan MI, Lei L. 2025 Learn then Test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics* **19**, 1641–1662.
37. Bates S, Angelopoulos A, Lei L, Malik J, Jordan M. 2021 Distribution-free, Risk-controlling Prediction Sets. *Journal of the ACM (JACM)* **68**, 1–34.
38. Gibbs I, Cherian JJ, Candès EJ. 2025 Conformal prediction with conditional guarantees. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **87**, 1100–1126.
39. Foygel Barber R, Candès EJ, Ramdas A, Tibshirani RJ. 2021 The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA* **10**, 455–482.
40. Vovk V. 2012 Conditional validity of inductive conformal predictors. In *Asian Conference on Machine Learning (ACML)* vol. 25 pp. 475–490. Proceedings of Machine learning Research.
41. Vovk V, Lindsay D, Nouretdinov I, Gammerman A et al.. 2003 Mondrian Confidence Machine. *Technical Report*.
42. Tibshirani RJ, Foygel Barber R, Candès E, Ramdas A. 2019 Conformal prediction under covariate shift. *Advances in Neural Information Processing Systems* **32**.
43. Barber RF, Candès EJ, Ramdas A, Tibshirani RJ. 2023 Conformal prediction beyond exchangeability. *Annals of Statistics* **51**, 816–845.
44. Romano Y, Patterson E, Candes E. 2019 Conformalized quantile regression. *Advances in Neural Information Processing Systems* **32**.
45. Vovk V. 2025 Conformal e-prediction. *Pattern Recognition* **166**, 111674.
46. Valiant LG. 1984 A theory of the learnable. *Communications of the ACM* **27**, 1134–1142.
47. Park S, Bastani O, Matni N, Lee I. 2020 PAC Confidence Sets for Deep Neural Networks via Calibrated Prediction. In *8th International Conference on Learning Representations (ICLR)*. https://openreview.net/forum?id=BJxVI04YvB.
48. Gibbs I, Candes E. 2021 Adaptive conformal inference under distribution shift. *Advances in Neural Information Processing Systems* **34**, 1660–1672. https://openreview.net/forum?id=6vaActvpcp3.
49. Deng Z, Zollo T, Snell J, Pitassi T, Zemel R. 2023 Distribution-free statistical dispersion control for societal applications. *Advances in Neural Information Processing Systems* **36**, 40342–40366.

50. 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 preprint arXiv:2305.18404*.
51. Ravfogel S, Goldberg Y, Goldberger J. 2023 Conformal nucleus sampling. In *Findings of the Association for Computational Linguistics: ACL 2023* pp. 27–34.
52. Ren AZ, Dixit A, Bodrova A, Singh S, Tu S, Brown N, Xu P, Takayama L, Xia F, Varley J et al.. 2023 Robots That Ask For Help: Uncertainty Alignment for Large Language Model Planners. In *Conference on Robot Learning* pp. 661–682. Proceedings of Machine Learning Research (PMLR).
53. Quach V, Fisch A, Schuster T, Yala A, Sohn JH, Jaakkola TS, Barzilay R. 2024 Conformal Language Modeling. In *12th International Conference on Learning Representations (ICLR)*. https://openreview.net/forum?id=pzUhfQ74c5.
54. Zollo TP, Morrill T, Deng Z, Snell J, Pitassi T, Zemel R. 2024 Prompt Risk Control: A Rigorous Framework for Responsible Deployment of Large Language Models. In *12th International Conference on Learning Representations (ICLR)*. https://openreview.net/forum?id=5tGGWOijvq.
55. Tsai YHH, Talbott W, Zhang J. 2024 Efficient non-parametric uncertainty quantification for black-box large language models and decision planning. *arXiv preprint arXiv:2402.00251*.
56. Ulmer D, Zerva C, Martins AF. 2024 Non-exchangeable conformal language generation with nearest neighbors. In *Findings of the Association for Computational Linguistics* pp. 1909–1929.
57. Deutschmann N, Alberts M, Martínez MR. 2024 Conformal autoregressive generation: Beam search with coverage guarantees. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence (AAAI)* vol. 38 pp. 11775–11783.
58. Rouzrokh P, Faghani S, Gamble CU, Shariatnia M, Erickson BJ. 2024 CONFLARE: Conformal large language model retrieval. *arXiv preprint arXiv:2404.04287*.
59. Yadkori YA, Kuzborskij I, Stutz D, György A, Fisch A, Doucet A, Beloshapka I, Weng WH, Yang YY, Szepesvári C et al.. 2024 Mitigating LLM hallucinations via conformal abstention. *arXiv preprint arXiv:2405.01563*.
60. Li S, Park S, Lee I, Bastani O. 2024 TRAQ: Trustworthy retrieval augmented question answering via conformal prediction. In *Proceedings of the 2024 Conference of the NA Chapter of the Association for Computational Linguistics: Human Language Technologies* pp. 3799–3821.
61. Kang M, Gürel NM, Yu N, Song D, Li B. 2024 C-RAG: Certified Generation Risks for Retrieval-Augmented Language Models. In *41st International Conference on Machine Learning (ICML)* vol. 235 pp. 22963–23000. Proceedings of Machine Learning Research (PMLR).
62. Ren AZ, Clark J, Dixit A, Itkina M, Majumdar A, Sadigh D. 2024 Explore until Confident: Efficient Exploration for Embodied Question Answering. In *1st Workshop on Vision-Language Models for Navigation and Manipulation at ICRA 2024*.
63. Giovannotti P, Gammerman A. 2024 Calibrated Large Language Models for Binary Question Answering. In *13th Symposium on Conformal and Probabilistic Prediction with Applications (COPA)* vol. 230 pp. 218–235. Proceedings of Machine Learning Research (PMLR).
64. Vovk V, Petej I. 2014 Venn-Abers predictors. In *Proceedings of the 13th Conference on Uncertainty in Artificial Intelligence (UAI)* pp. 829–838.
65. Mohri C, Hashimoto T. 2024 Language Models with Conformal Factuality Guarantees. In *41st International Conference on Machine Learning (ICML)* vol. 235 pp. 36029–36047. Proceedings of Machine Learning Research (PMLR).
66. Kaur R, Samplawski C, Cobb AD, Roy A, Matejek B, Acharya M, Elenius D, Berenbeim AM, Pavlik JA, Bastian ND, Jha S. 2024 Addressing Uncertainty in LLMs to Enhance Reliability in Generative AI. In *NeurIPS Safe Generative AI Workshop 2024*.
67. Rubin-Toles M, Gambhir M, Ramji K, Roth A, Goel S. 2025 Conformal Language Model Reasoning with Coherent Factuality. In *13th International Conference on Learning Representations (ICLR)*. https://openreview.net/forum?id=AJpUZd8Clb.
68. Frankel E, Li SS, Ratliff LJ, Tsvetkov Y, Oh S, Koh PW. 2025 Conformal Reasoning: Uncertainty Estimation in Interactive Environments. https://openreview.net/forum?id=Vf5ZUalFk8.
69. Wang Q, Geng T, Wang Z, Wang T, Fu B, Zheng F. 2025 Sample then Identify: A General Framework for Risk Control and Assessment in Multimodal Large Language Models. In *13th International Conference on Learning Representations (ICLR)*. https://openreview.net/forum?id=9WYMDgxDac.

70. Ni B. 2024 Reliable knowledge graph reasoning with uncertainty quantification. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM)* pp. 5463–5466.
71. Ni B, Wang Y, Cheng L, Blasch E, Derr T. 2025 Towards trustworthy knowledge graph reasoning: An uncertainty aware perspective. In *Proceedings of the 39th AAAI Conference on Artificial Intelligence (AAAI)* vol. 39 pp. 12417–12425.
72. Rahim N, Rahim AA. 2024 Probabilistic Proof State Compression: Optimizing LLM-Guided Formal Verification. In *4th Workshop on Mathematical Reasoning and AI at NeurIPS'24*.
73. Su J, Luo J, Wang H, Cheng L. 2024 API is enough: Conformal prediction for large language models without logit-access. In *Findings of the Association for Computational Linguistics: EMNLP 2024* pp. 979–995.
74. Huang J, Xi H, Zhang L, Yao H, Qiu Y, Wei H. 2024 Conformal Prediction for Deep Classifier via Label Ranking. In *41st International Conference on Machine Learning (ICML)* vol. 235 pp. 20331–20347. Proceedings of Machine Learning Research (PMLR).
75. Wang Z, Duan J, Cheng L, Zhang Y, Wang Q, Shi X, Xu K, Shen HT, Zhu X. 2024 CONU: Conformal uncertainty in large language models with correctness coverage guarantees. In *Findings of the Association for Computational Linguistics: EMNLP 2024* pp. 6886–6898.
76. Bai T, Jin Y. 2024 Optimized Conformal Selection: Powerful selective inference after conformity score optimization. *arXiv preprint arXiv:2411.17983*.
77. Intrator Y, Cohen R, Kelner O, Goldenberg R, Rivlin E, Freedman D. 2024 Streamlining conformal information retrieval via score refinement. In *7th Fact Extraction and VERification Workshop (FEVER)* pp. 186–191.
78. Wang S, Huang L. 2024 Debate as optimization: Adaptive conformal prediction and diverse retrieval for event extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2024* pp. 16422–16435.
79. Cherian JJ, Gibbs I, Candès EJ. 2024 Large language model validity via enhanced conformal prediction methods. *Advances in Neural Information Processing Systems* **37**, 114812–114842.
80. Ye F, Yang M, Pang J, Wang L, Wong DF, Yilmaz E, Shi S, Tu Z. 2024 Benchmarking LLMs via uncertainty quantification. *Advances in Neural Information Processing Systems* **37**, 15356–15385.
81. Gui Y, Jin Y, Ren Z. 2024 Conformal alignment: Knowing when to trust foundation models with guarantees. *Advances in Neural Information Processing Systems* **37**, 73884–73919.
82. Kiyani S, Pappas G, Hassani H. 2024 Length optimization in conformal prediction. *Advances in Neural Information Processing Systems* **37**, 99519–99563.
83. Liang K, Zhang Z, Fisac JF. 2024 Introspective planning: Aligning robots' uncertainty with inherent task ambiguity. *Advances in Neural Information Processing Systems* **37**, 71998–72031.
84. Lee M, Kim K, Kim T, Park S. 2024 Selective generation for controllable language models. *Advances in Neural Information Processing Systems* **37**, 50494–50527.
85. Wang J, He G, Kantaros Y. 2024 Probabilistically correct language-based multi-robot planning using conformal prediction. *IEEE Robotics and Automation Letters* **10**, 160–167.
86. Xu B, Lu Y. 2025 TECP: Token-Entropy Conformal Prediction for LLMs. *Mathematics* **13**, 3351.
87. Jung J, Brahman F, Choi Y. 2025 Trust or Escalate: LLM Judges with Provable Guarantees for Human Agreement. In *13th International Conference on Learning Representations (ICLR)*. https://openreview.net/forum?id=UHPnqSTBPO.
88. Tayebati S, Kumar D, Darabi N, Jayasuriya D, Krishnan R, Trivedi AR. 2025a Learning conformal abstention policies for adaptive risk management in large language and vision-language models. *arXiv preprint arXiv:2502.06884*.
89. Tayebati S, Kumar D, Darabi N, Jayasuriya D, Tulabandhula T, Krishnan R, Trivedi AR. 2025b CAP: Conformalized abstention policies for context-adaptive risk management for LLMs and VLMs. In *Proceedings of the 17th Asian Conference on Machine Learning (ACML)*.
90. Xi H, Huang J, Liu K, Feng L, Wei H. 2025 Does confidence calibration improve conformal prediction?. *Transactions on Machine Learning Research (TMLR)*. https://openreview.net/forum?id=6DDaTwTvdE.
91. Zhi Z, Feng C, Daneshmend A, Orlu M, Demosthenous A, Yin L, Li D, Liu Z, Rodrigues MR. 2025 Seeing and reasoning with confidence: Supercharging multimodal LLMs with an uncertainty-aware agentic framework. *arXiv preprint arXiv:2503.08308*.
92. Kaur N, McPheat L, Russo A, Cohn AG, Madhyastha P. 2025 An empirical study of conformal prediction in LLM with ASP scaffolds for robust reasoning. *arXiv preprint arXiv:2503.05439*.
93. Kim J, O'Hagan S, Rockova V. 2024 Adaptive uncertainty quantification for generative AI. *arXiv preprint arXiv:2408.08990*.

94. Ye Y, Wen W. 2025 Data-driven calibration of prediction sets in large vision-language models based on inductive conformal prediction. *arXiv preprint arXiv:2504.17671*.
95. Doula A, Mühlhäuser M, Guinea AS. 2025 SafePath: Conformal prediction for safe LLM-based autonomous navigation. *arXiv preprint arXiv:2505.09427*.
96. Sundarsingh DS, Wang J, Deshmukh JV, Kantaros Y. 2025 ConformalNL2LTL: Translating natural language instructions into temporal logic formulas with conformal correctness guarantees. *arXiv preprint arXiv:2504.21022*.
97. Ke Y, Lin H, Ruan Y, Tang J, Li L. 2025 Correctness Coverage Evaluation for Medical Multiple-Choice Question Answering Based on the Enhanced Conformal Prediction Framework. *Mathematics* **13**, 1538.
98. Kladny KR, Schölkopf B, Muehlebach M. 2025 Conformal Generative Modeling with Improved Sample Efficiency through Sequential Greedy Filtering. In *13th International Conference on Learning Representations (ICLR)*. https://openreview.net/forum?id=1i6lkavJ94.
99. Su J, Lin F, Feng Z, Zheng H, Wang T, Xiao Z, Zhao X, Liu Z, Cheng L, Wang H. 2026 CP-Router: An uncertainty-aware router between LLM and LRM. In *Proceedings of the 40th AAAI Conference on Artificial Intelligence (AAAI)* vol. 40 pp. 33065–33073.
100. Vishwakarma H, Mishler A, Cook T, Dalmasso N, Raman N, Ganesh S. 2025 Prune'n Predict: Optimizing LLM Decision-making with Conformal Prediction. In *42nd International Conference on Machine Learning (ICML)* pp. 61601–61634. Proceedings of Machine Learning Research.
101. Shahrokhi H, Roy DR, Yan Y, Arnaoudova V, Doppa J. 2025 Conformal Prediction Sets for Deep Generative Models via Reduction to Conformal Regression. In *41st Conference on Uncertainty in Artificial Intelligence* pp. 3718–3748. Proceedings of Machine Learning Research.
102. Wang Z, Duan J, Wang Q, Zhu X, Chen T, Shi X, Xu K. 2026 COIN: Uncertainty-guarding selective question answering for foundation models with provable risk guarantees. In *Proceedings of the 40th AAAI Conference on Artificial Intelligence (AAAI)* vol. 40 pp. 33764–33772.
103. Chen C, Shen J, Deng Z, Lei L. 2025 Conformal Tail Risk Control for Large Language Model Alignment. In *42nd International Conference on Machine Learning (ICML)* vol. 267 pp. 8955–8978. Proceedings of Machine Learning Research (PMLR).
104. Nag S, Ghosh U, Ta CK, Bose S, Li J, Roy-Chowdhury AK. 2025 Conformal prediction and MLLM aided uncertainty quantification in scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* pp. 11676–11686.
105. Fayyazi A, Kamal M, Pedram M. 2025 FACTER: Fairness-Aware Conformal Thresholding and Prompt Engineering for Enabling Fair LLM-Based Recommender Systems. In *42nd International Conference on Machine Learning (ICML)* vol. 267 pp. 16407–16422. Proceedings of Machine Learning Research (PMLR).
106. Noorani S, Kiyani S, Pappas GJ, Hassani H. 2025 Conformal Prediction Beyond the Seen: A Missing Mass Perspective for Uncertainty Quantification in Generative Models. In *ICML 2025 Workshop on Reliable and Responsible Foundation Models*. https://openreview.net/forum?id=vGg1Nvd4xo.
107. Chan KHR, Ge Y, Dobriban E, Hassani H, Vidal R. 2025 Conformal Information Pursuit for Interactively Guiding Large Language Models. In *Proceedings of the 39th Annual Conference on Neural Information Processing Systems (NeurIPS)*. https://openreview.net/forum?id=xAHozxfuUW.
108. Wang Z, Wang Q, Zhang Y, Chen T, Zhu X, Shi X, Xu K. 2025 SConU: Selective conformal uncertainty in large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)* pp. 19052–19075.
109. Gui Y, Jin Y, Nair Y, Ren Z. 2025 ACS: An interactive framework for conformal selection. *arXiv preprint arXiv:2507.15825*.
110. Liu T, Wu ZS. 2025 Multi-group uncertainty quantification for long-form text generation. In *Proceedings of the 41st Conference on Uncertainty in Artificial Intelligence (UAI)* pp. 2659–2684.
111. Feng N, Sui Y, Hou S, Cresswell JC, Wu G. 2025 Response quality assessment for retrieval-augmented generation via conditional conformal factuality. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval* pp. 2832–2836.
112. Chen Y, Yin CH, Chikodikar SM, Vinayak RK. 2025 On the Scoring Functions for RAG-based Conformal Factuality. In *ICML 2025 Workshop on Reliable and Responsible Foundation Models*.
113. Xie Z, Hong Z, Lyu W, Wang H, Wang G, Zhang D. 2025 CoAlign: Uncertainty Calibration

of LLM for Geospatial Repartition. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)* pp. 254–263.
114. Ye Y. 2025 Conformal P-Value in Multiple-Choice Question Answering Tasks with Provable Risk Control. *arXiv preprint arXiv:2508.10022*.
115. Liu K, Xi H, Vong CM, Wei H. 2026 Online conformal selection with accept-to-reject changes. In *Proceedings of the 40th AAAI Conference on Artificial Intelligence* vol. 40 pp. 23765–23773.
116. Zhang L, Wang C. 2025 Selective Prediction for VQA: Enhancing Trust in MLLMs Through Normalized Edit-Distance Conformal Calibration. In *Proceedings of the 6th International Conference on Artificial Intelligence and Electromechanical Automation (AIEA)* pp. 01–05. IEEE.
117. Zhao H, Zhu Y, Wang Z, Wang Y, Gao J, Ma L. 2025 ConfAgents: A conformal-guided multi-agent framework for cost-efficient medical diagnosis. *arXiv preprint arXiv:2508.04915*.
118. Yang G, Liu X. 2025 Conformal Sets in Multiple-Choice Question Answering under Black-Box Settings with Provable Coverage Guarantees. *arXiv preprint arXiv:2508.05544*.
119. Yang G, Zhang Y, Liu X, Wu Z. 2025 Frequency-Based Predictive Entropy for Uncertainty Quantification in Black-Box Multiple-Choice Question Answering. In *Proceedings of the 4th International Conference on Image Processing, Computer Vision and Machine Learning (ICICML)* pp. 1782–1786. IEEE.
120. Wang J, Tong J, Tan K, Vorobeychik Y, Kantaros Y. 2025 Conformal temporal logic planning using large language models. *ACM Transactions on Cyber-Physical Systems*.
121. Wu Y, Wu S, Tao Y, Li Y, Sarwate AD. 2025 Not only a helper, but also a teacher: Interactive LLM Cascade. *arXiv preprint arXiv:2509.22984*.
122. Oehri M, Conti G, Pather K, Rossi A, Serra L, Parody A, Johannesen R, Petersen A, Krasniqi A. 2025 Trusted Uncertainty in Large Language Models: A Unified Framework for Confidence Calibration and Risk-Controlled Refusal. *arXiv preprint arXiv:2509.01455*.
123. Wu Z, Jeong SW, Liu Y, Jung YJ, Donnat C. 2025 Filtering with Confidence: When Data Augmentation Meets Conformal Prediction. *arXiv preprint arXiv:2509.21479*.
124. Kostyumov V, Nutfullin B, Pilipenko O. 2025 Evaluation of Multimodal Image and Text Processing Models from an Uncertainty Perspective. *Pattern Recognition and Image Analysis* **35**, 255–268.
125. Chen Y, Wang Y, Liu S, Jing Y, Tao D. 2025 CoVeR: Conformal Calibration for Versatile and Reliable Autoregressive Next-Token Prediction. *arXiv preprint arXiv:2509.04733*.
126. Azad A, Hossain MS, Shanto MSH, Rahman MS, Parvez MR. 2026 The Art of Saying "Maybe": A Conformal Lens for Uncertainty Benchmarking in VLMs. In *Findings of the Association for Computational Linguistics: EACL 2026* pp. 5185–5201.
127. Pang L, Huang L, Lin J, Wang T, Horiguchi A, Aue A, Priebe CE. 2025a Unsupervised Conformal Inference: Bootstrapping and Alignment to Control LLM Uncertainty. *arXiv preprint arXiv:2509.23002*.
128. Pang L, Huang L, Lin J, Wang T, Aue A, Priebe CE. 2025b Taming Variability: Randomized and Bootstrapped Conformal Risk Control for LLMs. *arXiv preprint arXiv:2509.23007*.
129. Wang NS, Yaldiz DN, Bakman YF, Karimireddy SP. 2025a Conformal Prediction Adaptive to Unknown Subpopulation Shifts. *arXiv preprint arXiv:2506.05583*.
130. Wang X, Hu G, Peng L, Zou C. 2025b AggLCF: Aggregation Enhanced Localized Conformal Factuality for Large Language Models. .
131. Overman W, Bayati M. 2025 Conformal Arbitrage: Risk-Controlled Balancing of Competing Objectives in Language Models. In *Proceedings of the 39th Annual Conference on Neural Information Processing Systems (NeurIPS)*. https://openreview.net/forum?id=dX2BTCD02T.
132. Lin Z, Li Y, Sarna N, Gao Y, von Gablenz M. 2025 Domain-shift-aware conformal prediction for large language models. *arXiv preprint arXiv:2510.05566*.
133. Wang Q, Fan Y, Wang XE. 2026 SAFER: Risk-Constrained Sample-then-Filter in Large Language Models. In *14th International Conference on Learning Representations (ICLR)*. https://openreview.net/forum?id=kJmLmOvwLC.
134. Jiang Z, Liu A, Durme BV. 2025 Conformal Linguistic Calibration: Trading-off between Factuality and Specificity. In *Proceedings of the 39th Annual Conference on Neural Information Processing Systems (NeurIPS)*. https://openreview.net/forum?id=MWF1ZzYnxJ.
135. Zhou X, Cheng L. 2025 Robust Uncertainty Quantification for Self-Evolving Large Language Models via Continual Domain Pretraining. *arXiv preprint arXiv:2510.22931*.
136. Dhillon GS, González J, Pandeva T, Curth A. 2025 E-Scores for (In) Correctness Assessment of Generative Model Outputs. *arXiv preprint arXiv:2510.25770*.

137. Hu Z, Zhang Z, Wang Y, Rutkowski L, Tao D. 2026 CoFact: Conformal Factuality Guarantees for Language Models under Distribution Shift. In *14th International Conference on Learning Representations (ICLR)*. https://openreview.net/forum?id=eiBp7rsc3K.
138. Xiong J, Chen Q, Ye F, Wan Z, Zheng C, Zhao C, Shen H, Li AH, Tao C, Tan H, Bai H, Shang L, Kong L, Wong N. 2026 ATTS: Asynchronous Test-Time Scaling via Conformal Prediction. In *14th International Conference on Learning Representations (ICLR)*. https://openreview.net/forum?id=YM3SskmtCE.
139. Hao Q, Liao W, Jing B, Wei H. 2026 Multi-Condition Conformal Selection. In *14th International Conference on Learning Representations (ICLR)*. https://openreview.net/forum?id=giL8Q1V26J.
140. Ashok D, May J. 2025 Language Models Can Predict Their Own Behavior. In *Proceedings of the 39th Annual Conference on Neural Information Processing Systems (NeurIPS)*. https://openreview.net/forum?id=i8IqEzpHaJ.
141. Xin J, Qiang E, Li X, Su WJ, Long Q. 2026 Paraphrase-Robust Conformal Prediction for Reliable LLM Uncertainty Quantification. .
142. Xue N, Chen Z, Yao J, Sun Y, Tao M. 2026 Conformal Risk-Controlled Routing for Large Language Model. .
143. Chen Y, Chen D, Chikodikar SM, Yin CH, Vinayak RK. 2026 Understanding Conformal Factuality for RAG-based LLMs: Novel Metrics and Systematic Insights. .
144. Bai T, Zhao Y, Yu X, Yang AY. 2025 Multivariate Conformal Selection. In *42nd International Conference on Machine Learning (ICML)* vol. 267 pp. 2535–2559. Proceedings of Machine Learning Research (PMLR).
145. Wang S, Jiang Y, Tang Y, Cheng L, Chen H. 2025 COPU: Conformal prediction for uncertainty quantification in natural language generation. *arXiv preprint arXiv:2502.12601*.
146. Wang T, Sun Y, Dobriban E. 2026 Singleton-Optimized Conformal Prediction. In *14th International Conference on Learning Representations (ICLR)*. https://openreview.net/forum?id=mO3nEGibLA.
147. Kuwahara B, Lin CY, Huang XS, Leung KK, Yapeter JA, Stanevich I, Perez F, Cresswell JC. 2025 Document Summarization with Conformal Importance Guarantees. In *Proceedings of the 39th Annual Conference on Neural Information Processing Systems (NeurIPS)*. https://openreview.net/forum?id=w1Y7RZC3QT.
148. Noh K, Lee S, Kim I, Song K. 2026 Multi-LLM Adaptive Conformal Inference for Reliable LLM Response. In *14th International Conference on Learning Representations (ICLR)*. https://openreview.net/forum?id=opuQH9Xyu9.
149. Sheng H, Liu X, He H, Zhao J, Kang J. 2025 Analyzing Uncertainty of LLM-as-a-Judge: Interval Evaluations with Conformal Prediction. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)* pp. 11297–11339.
150. Sesia M, Candès EJ. 2020 A comparison of some conformal quantile regression methods. *Stat* **9**, e261.
151. Sesia M, Romano Y. 2021 Conformal prediction using conditional histograms. *Advances in Neural Information Processing Systems* **34**, 6304–6315.
152. Lin Z, Trivedi S, Sun J. 2021 Locally valid and discriminative prediction intervals for deep learning models. *Advances in Neural Information Processing Systems* **34**, 8378–8391.
153. Xie R, Barber RF, Candès EJ. 2024 Boosted conformal prediction intervals. *Advances in Neural Information Processing Systems* **37**, 71868–71899.
154. Guha EK, Natarajan S, Möllenhoff T, Khan ME, Ndiaye E. 2024 Conformal Prediction via Regression-as-Classification. In *12th International Conference on Learning Representations (ICLR)*. https://openreview.net/forum?id=rulxyXjf46.
155. Lu C, Angelopoulos AN, Pomerantz S. 2022 Improving trustworthiness of AI disease severity rating in medical imaging with ordinal conformal prediction sets. In *International Conference on Medical Image Computing and Computer-assisted Intervention* pp. 545–554. Springer.
156. Xu Y, Guo W, Wei Z. 2023 Conformal risk control for ordinal classification. In *39th Conference on Uncertainty in Artificial Intelligence* vol. 216 pp. 2346–2355. Proceedings of Machine learning Research (PMLR).
157. Li Z, Yan C, Jackson NJ, Cui W, Li B, Zhang J, Malin BA. 2025 Towards statistical factuality guarantee for large vision-language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)* pp. 11446–11467.
158. Gupta A, Kaur R, Roy A, Cobb AD, Chellappa R, Jha S. 2025 Polysemantic Dropout:

Conformal OOD Detection for Specialized LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)* pp. 11768–11781.
159. Wang J, Vorobeychik Y, Kantaros Y. 2025 CoFineLLM: Conformal Finetuning of LLMs for Language-Instructed Robot Planning. *arXiv preprint arXiv:2511.06575*.
160. Zhang T, Bisht N, Li Z, Xu G, Wang X. 2025 SarRec: Statistically-guaranteed Augmented Retrieval for Recommendation. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management (CIKM)* pp. 4181–4190.
161. Kim E, Foty R, Shrestha M, Seyfert-Margolis V. 2025 Conformal prediction and verification of large language model extractions in EHR data. In *Proceedings of the AAAI Symposium Series* vol. 7 pp. 539–546.
162. Li H, He Z, Chen X, Zhang C, Quan SF, Killgore WD, Wung SF, Chen CX, Yuan G, Lu J et al.. 2025 Smarter Together: Combining Large Language Models and Small Models for Physiological Signals Visual Inspection. *Journal of Healthcare Informatics Research* **9**, 656–685.
163. Hao S, Zeng H, Wei H, Jing B. 2026 RACER: Risk-Aware Calibrated Efficient Routing for Large Language Models. *arXiv preprint arXiv:2603.06616*.
164. Ye K, Pan Q, Li S. 2026 Conditional Factuality Controlled LLMs with Generalization Certificates via Conformal Sampling. *arXiv preprint arXiv:2603.27403*.
165. Chakraborty D, Yang E, Khashabi D, Lawrie D, Duh K. 2026 Principled Context Engineering for RAG: Statistical Guarantees via Conformal Prediction. In *Proceedings of the 48th European Conference on Information Retrieval (ECIR)* pp. 537–546. Springer.
166. Vovk V, Fedorova V, Nouretdinov I, Gammerman A. 2016 Criteria of efficiency for conformal prediction. In *5th Symposium on Conformal and Probabilistic Prediction with Applications (COPA)* pp. 23–39. Springer.
167. Hendrycks D, Burns C, Basart S, Zou A, Mazeika M, Song D, Steinhardt J. 2021 Measuring Massive Multitask Language Understanding. In *9th International Conference on Learning Representations (ICLR)*. https://openreview.net/forum?id=d7KBjmI3GmQ.
168. Pal A, Umapathi LK, Sankarasubbu M. 2022 MedMCQA: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the 2nd Conference on Health, Inference, and Learning (CHIL)* vol. 174 pp. 248–260. Proceedings of Machine Learning Research (PMLR).
169. Joshi M, Choi E, Weld DS, Zettlemoyer L. 2017 TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)* pp. 1601–1611.
170. Reddy S, Chen D, Manning CD. 2019 CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics (TACL)* **7**, 249–266.
171. Kwiatkowski T, Palomaki J, Redfield O, Collins M, Parikh A, Alberti C, Epstein D, Polosukhin I, Devlin J, Lee K et al.. 2019 Natural Questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics* **7**, 453–466.
172. Cobbe K, Kosaraju V, Bavarian M, Chen M, Jun H, Kaiser L, Plappert M, Tworek J, Hilton J, Nakano R et al.. 2021 Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.
173. Hendrycks D, Burns C, Kadavath S, Arora A, Basart S, Tang E, Song D, Steinhardt J. 2021 Measuring Mathematical Problem Solving With the MATH Dataset. In *Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*.
174. Yue X, Ni Y, Zhang K, Zheng T, Liu R, Zhang G, Stevens S, Jiang D, Ren W, Sun Y et al.. 2024 MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* pp. 9556–9567.
175. Kembhavi A, Salvato M, Kolve E, Seo M, Hajishirzi H, Farhadi A. 2016 A diagram is worth a dozen images. In *Proceedings of the 14th European Conference on Computer Vision (ECCV)* pp. 235–251. Springer.
176. Lu P, Mishra S, Xia T, Qiu L, Chang KW, Zhu SC, Tafjord O, Clark P, Kalyan A. 2022 Learn to Explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems* **35**, 2507–2521.
177. Johnson AE, Pollard TJ, Berkowitz SJ, Greenbaum NR, Lungren MP, Deng Cy, Mark RG, Horng S. 2019 MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data* **6**, 317.
178. Min S, Krishna K, Lyu X, Lewis M, Yih Wt, Koh P, Iyyer M, Zettlemoyer L, Hajishirzi H. 2023 FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In

Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP) pp. 12076–12100.
179. Jeong M, Hwang H, Yoon C, Lee T, Kang J. 2024 OLAPH: Improving factuality in biomedical long-form question answering. *arXiv preprint arXiv:2405.12701*.
180. Touvron H, Lavril T, Izacard G, Martinet X, Lachaux MA, Lacroix T, Rozière B, Goyal N, Hambro E, Azhar F et al.. 2023a LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* **10**.
181. Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y, Bashlykov N, Batra S, Bhargava P, Bhosale S et al.. 2023b LLaMA 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
182. Grattafiori A, Dubey A, Jauhri A, Pandey A, Kadian A, Al-Dahle A, Letman A, Mathur A, Schelten A, Vaughan A et al.. 2024 The LLaMA 3 herd of models. *arXiv preprint arXiv:2407.21783*.
183. Bai J, Bai S, Chu Y, Cui Z, Dang K, Deng X, Fan Y, Ge W, Han Y, Huang F et al.. 2023 Qwen technical report. *arXiv preprint arXiv:2309.16609*.
184. Yang A, Yang B, Hui B, Zheng B, Yu B, Zhou C, Li C, Li C, Liu D, Huang F et al.. 2024a Qwen2 technical report. *eprint arXiv: 2407.10671*.
185. Yang A, Yang B, Zhang B, Hui B, Zheng B, Yu B, Li C, Liu D, Huang F, Wei H et al.. 2024b Qwen2.5 technical teport. *arXiv preprint arXiv:2412.15115*.
186. Yang A, Li A, Yang B, Zhang B, Hui B, Zheng B, Yu B, Gao C, Huang C, Lv C et al.. 2025 Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.
187. Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman FL, Almeida D, Altenschmidt J, Altman S, Anadkat S et al.. 2023 GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.
188. Kirillov A, Jiang A, Rossen B, Bassin C, Hudson C, Shern CJ, Fischer C, Sherburn D, Mays E, Raso F, von Lohmann F, Sulit F, Starace G, Aung J, Lennon J, Phang J, Lee JG, Candela JQ, Parish J, Uesato J, Singhal K, Shi K, Wood K, Liu K et al.. 2024 GPT-4o System Card. Technical report OpenAI. https://openai.com/index/gpt-4o-system-card/.
189. Jiang AQ, Sablayrolles A, Mensch A, Bamford C, Singh Chaplot D, de las Casas D, Bressand F, Lengyel G, Lample G, Saulnier L, Renard Lavaud L, Lachaux MA, Stock P, le Scao T, Lavril T, Wang T, Lacroix T, el Sayed W. 2023 Mistral 7B. *arXiv preprint arXiv:2310.06825*.
190. Jiang AQ, Sablayrolles A, Roux A, Mensch A, Savary B, Bamford C, Chaplot DS, Casas Ddl, Hanna EB, Bressand F et al.. 2024 Mixtral of Experts. *arXiv preprint arXiv:2401.04088*.
191. Liu A, Feng B, Xue B, Wang B, Wu B, Lu C, Zhao C, Deng C, Zhang C, Ruan C et al.. 2024 DeepSeek-R3 technical report. *arXiv preprint arXiv:2412.19437*.
192. Guo D, Yang D, Zhang H, Song J, Wang P, Zhu Q, Xu R, Zhang R, Ma S, Bi X et al.. 2025 DeepSeek-R1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.
193. Team G, Anil R, Borgeaud S, Alayrac JB, Yu J, Soricut R, Schalkwyk J, Dai AM, Hauth A, Millican K et al.. 2023 Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.
194. Team G, Georgiev P, Lei VI, Burnell R, Bai L, Gulati A, Tanzer G, Vincent D, Pan Z, Wang S et al.. 2024 Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.
195. Team G, Comanici G, Bieber E, Schaekermann M, Pasupat I, Sachdeva N, Dhillon I, Blistein M, Ram O, Zhang D, Rosen E et al.. 2025 Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.
196. Chatterji A, Cunningham T, Deming D, Hitzig Z, Ong C, Shan C, Wadman K. 2025 How People Use ChatGPT. Technical report OpenAI. Accessed: 16/04/2026.
197. Murray C. 2026 Anthropic closes in on OpenAI as US business use surges. Accessed: 16/04/2026.
198. Morris S. 2026 Google set to double AI spending to $185bn after strong earnings. Accessed: 16/04/2026.
199. Tully T, Redfern J, Das D, Xiao D. 2025 The State of Generative AI in the Enterprise. Accessed: 03/04/2026.
200. Jin D, Pan E, Oufattole N, Weng WH, 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**, 6421.

201. Huang L, Le Bras R, Bhagavatula C, Choi Y. 2019 CosmosQA: Machine reading comprehension with contextual commonsense reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* pp. 2391–2401.
202. Talmor A, Herzig J, Lourie N, Berant J. 2019 CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North America Chapter of the Association for Computational Linguistics: Human Language Technologies* pp. 4149–4158.
203. Liu J, Cui L, Liu H, Huang D, Wang Y, Zhang Y. 2021 LogiQA: a challenge dataset for machine reading comprehension with logical reasoning. In *Proceedings of the 29th International Conference on International Joint Conferences on Artificial Intelligence* pp. 3622–3628.
204. Nimo C, Olatunji T, Owodunni AT, Abdullahi T, Ayodele E, Sanni M, Aka EC, Omofoye F, Yuehgoh F, Faniran T et al.. 2025 AfriMed-QA: A Pan-African, multi-specialty, medical question-answering benchmark dataset. In *The 63rd Annual Meeting of the Association for Computational Linguistics (ACL)* pp. 1948–1973.
205. Mallen A, Asai A, Zhong V, Das R, Khashabi D, Hajishirzi H. 2023 When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)* pp. 9802–9822.
206. Yang Z, Qi P, Zhang S, Bengio Y, Cohen W, Salakhutdinov R, Manning CD. 2018 HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)* pp. 2369–2380.
207. Min S, Michael J, Hajishirzi H, Zettlemoyer L. 2020 AmbigQA: Answering ambiguous open-domain questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* pp. 5783–5797.
208. Malaviya C, Lee S, Chen S, Sieber E, Yatskar M, Roth D. 2024 ExpertQA: Expert-curated questions and attributed answers. In *Proceedings of the 2024 Conference of the NA Chapter of the Association for Computational Linguistics: Human Language Technologies* pp. 3025–3045.
209. 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 (ACL)* pp. 3214–3252.
210. Bai Z, Wang P, Xiao T, He T, Han Z, Zhang Z, Shou MZ. 2024 Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*.
211. Yang T, Li Z, Cao J, Xu C. 2025 Understanding and mitigating hallucination in large vision-language models via modular attribution and intervention. In *13th International Conference on Learning Representations (ICLR)*. https://openreview.net/forum?id=Bjq4W7P2Us.
212. Wang X, Nalisnick E. 2026 Are vision language models robust to classic uncertainty challenges?. *Transactions on Machine Learning Research (TMLR)*. https://openreview.net/forum?id=4lCSYCNfmo.
213. Stutz D, Dvijotham KD, Cemgil AT, Doucet A. 2022 Learning Optimal Conformal Classifiers. In *10th International Conference on Learning Representations (ICLR)*. https://openreview.net/forum?id=t8O-4LKFVx.
214. Chen T, Guestrin C. 2016 XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* pp. 785–794.
215. Zhang H, Zhang Y, Yu Y, Madeka D, Foster D, Xing E, Lakkaraju H, Kakade S. 2024a A study on the calibration of in-context learning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* pp. 6118–6136.
216. Zhang M, Huang M, Shi R, Guo L, Peng C, Yan P, Zhou Y, Qiu X. 2024b Calibrating the confidence of large language models by eliciting fidelity. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)* pp. 2959–2979.
217. Plaut B, Nguyen KX, Trinh T. 2025 Probabilities of Chat LLMs Are Miscalibrated but Still Predict Correctness on Multiple-Choice Q&A. *Transactions on Machine Learning Research (TMLR)*. https://openreview.net/forum?id=E6LOh5vz5x.
218. Zhang YJ, Zhang ZY, Zhao P, Sugiyama M. 2023 Adapting to continuous covariate shift via online density ratio estimation. *Advances in Neural Information Processing Systems* **36**, 29074–29113. https://openreview.net/forum?id=ad3JNoR2np.

+