Conformal Language Modeling via Posterior Sampling

Nicolas Emmenegger* **Theo X. Olausson*** **Armando Solar-Lezama** **Chara Podimata**
Massachusetts Institute of Technology
{nemm,theoxo,asolar,podimata}@mit.edu

Abstract

Language models remain plagued by hallucinations. Recent work has sought to tame their prevalence using statistical techniques based on conformal prediction, with both theoretical and empirical success. However, these methods typically operate in a post-hoc fashion, treating the sampling procedure itself as atomic and then surgically altering samples to remove hallucinated claims. This disconnect between filtering and generation can result in samples that are incoherent, inconsistent, or simply unlikely under the model itself. Moreover, post-hoc surgery is unable to shift probability mass towards more useful and helpful responses. To address these issues, we propose to instead sample from approximations to a posterior distribution, where the conditioning event corresponds to a calibrated, high-scoring region. We develop a calibration procedure tailored to the setting of conditional sequential generation that effectively identifies this region and achieves target risk control. Empirically, we apply our method to case studies focused on open-ended biography generation and mathematical problem solving; compared to prior work, we obtain the same statistical guarantees, with higher downstream utility.

1 Introduction

Large Language Models (LLMs) are increasingly deployed in high-stakes settings ranging from medical diagnosis [Gaber et al., 2025] to legal research and case analysis [Dehghani et al., 2025]. Such settings require language models to act reliably, providing factual and coherent responses, yet even the most sophisticated models available today still routinely hallucinate false claims [Fan et al., 2026]. This has motivated a growing body of work on quantifying and reducing hallucinations, including methods that train LLMs to predict their own confidence [Damani et al., 2026], methods that produce calibrated predictions of factuality [Detommaso et al., 2024], or methods that generate sets of outputs containing a factual response with high probability [Quach et al., 2024].

One particularly promising recent direction has been to leverage *conformal prediction* [Mohri and Hashimoto, 2024, Cherian et al., 2024, Rubin-Toles et al., 2025], a statistical wrapper yielding assumption-lean guarantees. This line of work seeks to control the expected value of a loss function $\ell : X \times Y \to \mathbb{R}_{\geq 0}$, such as a measure of an output y's factuality with respect to a prompt x. To control this quantity, conformal methods split the output into distinct claims, and leverage a heuristic score function, which serves as an (imperfect) proxy of the loss. By observing the relationship between the scores and the true losses on a held-out calibration set, these methods then calibrate a *post-hoc filtering* mechanism that provably controls the risk. Formally, given a calibration set $D_{\text{cal}} = \{(x_i, y_i, \boldsymbol{s}_i, \ell_i)\}_{i=1}^n$ of prompt-response pairs with claim-level scores s_{ij} and observed losses $\ell_i = \ell(x_i, y_i)$, they construct a mapping F that filters out low-confidence claims so that the resulting $F(y_{\text{test}})$ satisfies $\mathbb{E}[\ell(x_{\text{test}}, F(y_{\text{test}}))] \leq \alpha$, for some user-defined risk tolerance $\alpha \in (0, 1)$.

While post-hoc surgical alteration of outputs makes sense in situations where the user is satisfied with a partially filtered generation, it may lead to samples that have low downstream utility in tasks where

*Equal contribution.
Our code and data are available at github.com/theoxo/posterior-conformal-lms.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

claims have strong interdependencies. In other words, while post-hoc filtering guarantees a bound on the probability that a response contains a false claim, this guarantee often comes at the cost of rendering many of the model's responses *incomplete* or simply *unhelpful* to the user.

To produce samples that are fluent, coherent and satisfying to the user, such methods therefore typically repair the filtered output using either a language model [Mohri and Hashimoto, 2024], thus reintroducing the risk of fabricating claims, or change the filtering mechanism symbolically through complicated dependency-graph schemes [Rubin-Toles et al., 2025]. However, even with some degree of fluency restored, these methods cannot recover from (over-)aggressively filtered responses without invalidating their theoretical guarantees .

Against this backdrop, in this paper, we ask: *Can we design a **risk-controlling** calibration routine that respects the underlying generative process and moves probability mass towards **useful** responses whenever the model has the capacity to solve the task at hand?*

We answer this question affirmatively by shifting the target of calibration from a post-hoc filter to the sampling distribution itself. Our key idea is that risk control should reweight the model's distribution towards completions that are more reliable and more useful, rather than altering the completions in a post-hoc manner. We reframe conformal language modeling as posterior sampling, conditioned on an event that corresponds to the response falling within a high-confidence region of the output space. That is, instead of calibrating a post-hoc filtering mechanism, we directly calibrate a sampling distribution $Q(\cdot|x_{\text{test}}, D_{\text{cal}})$ such that

$$\mathbb{E}_{D_{\text{cal}}, x_{\text{test}}, y_{\text{test}} \sim Q(\cdot|x_{\text{test}}, D_{\text{cal}})}[\ell(x_{\text{test}}, y_{\text{test}})] \leq \alpha. \tag{1}$$

1.1 Roadmap and Contributions

Modeling. In Sec. 2, we formulate conformal factuality control as *posterior sampling* from a calibrated family, indexed by a one-dimensional parameter τ. We define a family of potentials which threshold completions by their cumulative claim-level factuality scores. This amounts to conditioning the language model's distribution on high-confidence regions. While, as we show, this is cleanly motivated by a probabilistic formulation, our method's safety guarantees are agnostic to the choice of score function.

Calibration. In Sec. 3, we develop an *off-policy* calibration procedure for selecting the posterior threshold τ, *without requiring samples from each candidate posterior*. We face two major challenges: firstly, the posterior's conditional normalizing constant $Z^\tau(x)$ now needs to be estimated from samples, which is ill-conditioned (or even ill-defined) when the model has low (or no) mass on the high-confidence region. Secondly, estimating a plug-in of the importance reweighted loss naively yields a non-monotone objective, making conformal guarantees difficult to obtain. We solve the former issue by modeling the calibration objective as an *empirical distribution posterior* that is extended by an explicit abstention path, effectively modeling the prior as a mixture distribution. We solve the second issue by explicitly monotonizing the objective and experimentally validating that this has a negligible impact on conservativeness. Finally, we connect the empirical posterior distribution to the true population posterior; using a Glivenko-Cantelli argument, we show that the empirical posterior risk converges to the population posterior risk as the number of calibration samples grows.

Experiments. In Sec. 4, we evaluate our method on two case studies: open-ended biography generation [Min et al., 2023] and math reasoning [Hendrycks et al., 2021]. Our method tracks the desired factuality level while significantly *improving downstream utility* relative to a post-hoc conformal filtering baseline, especially in settings with strong claim interdependencies.

2 Preliminaries and Modeling

2.1 The Posterior Family

We model a language model with vocabulary $\mathcal{V}$ and prompt x as a distribution $p_{\text{LLM}}(\cdot \mid x) \in \Delta(\mathcal{V}^*)$ over completions, where $\mathcal{V}^*$ is the Kleene closure of $\mathcal{V}$ (up to some max length L, omitted for brevity). We call p_{LLM} the LM *prior* and extend it with a potential function $\Phi(y \mid x)$ to obtain a *posterior*

$$p^{\Phi}_{\text{LLM}}(y \mid x) \propto p_{\text{LLM}}(y \mid x)\Phi(y \mid x). \tag{2}$$

Our goal is to pick Φ so that substituting $p^{\Phi}_{\mathrm{LLM}}(\cdot|x)$ for $Q(\cdot|x)$ in Equation (1) controls risk at the target level. To do so, we make $\Phi(\cdot|x)$ depend on a heuristic notion of how "risky" y is, and control the perceived risk by setting a one-dimensional parameter τ, yielding a posterior $p^{\Phi^\tau}_{\mathrm{LLM}}(\cdot|x) = p^{\tau}_{\mathrm{LLM}}(\cdot|x)$ defined via the parametrized potentials Φ^τ. The right-hand side of Equation (2) implicitly depends on the *normalization constant* $Z^\tau(x) = \int_{y\in\mathcal{V}^*} \Phi^\tau(y|x)\,\mathrm{d}p_{\mathrm{LLM}}(y|x)$, which—as we will see below—corresponds to the probability mass that survives the threshold τ for prompt x. We next turn to the question of how to instantiate Φ^τ for a given τ.

2.2 From Heuristic Scores to Potentials

To build a family of potential functions that allow us to control the risk, we draw inspiration from prior work on conformal prediction and leverage a heuristic scoring function. We design this scoring function so that it gives an estimated confidence for y's most recent claim c_i being correct conditional on the prompt x and all preceding claims $\boldsymbol{c}_{<i} = (c_1, \ldots, c_{i-1})$ being assumed to be correct.[2] Formally, that is, the scoring function is of the form $s : \mathcal{V}^* \times \mathcal{V}^* \times (\mathcal{V}^*)^* \to \mathcal{S}$, where $\mathcal{S} \subset \mathbb{R}$ is chosen to be $(0, 1]$ w.l.o.g. Using $\mathbf{c}_{<i}$ to denote the prior claims whenever clear from context, we will abuse probabilistic conditioning notation and write $s(c_i, x, (c_1, \ldots, c_{i-1}))$ as

$$s_i := s(c_i \,|\, x, \mathbf{c}_{<i}) := s(c_i \,|\, x, c_1, \ldots, c_{i-1}).$$

Intuitively, the "oracle" score

$$s(c_i \,|\, x, \boldsymbol{c}_{<i}) = \mathbf{1}\left\{c_i \text{ is factual given } \boldsymbol{c}_{<i} \text{ and prompt } x\right\}$$

would immediately give us an $\alpha = 0$ level risk-controlling potential by letting $\Phi(y \,|\, x) = \mathbf{1}\left\{\prod_i s_i = 1\right\}$, ensuring that all samples from p^{Φ}_{LLM} contain only true claims. In practice, such an oracle cannot be obtained. The purpose of this paper is to show that with a heuristic choice of s, we can, however, still control the risk by appropriately picking a threshold τ such that

$$\Phi^\tau(y \,|\, x) = \mathbf{1}\left\{\prod_{i=1}^{K} s(c_i \,|\, x, \mathbf{c}_{<i}) > \tau\right\}, \tag{3}$$

where K is the (random) number of claims in the output. We will sometimes write $S(y \,|\, x) = \prod_{i\in[K]} s(c_i \,|\, x, \boldsymbol{c}_{<i})$ to denote completion-level scores. We build intuition for this definition by recalling the chain rule of probabilities. Assume for simplicity that we are in an idealized scenario where the scores s_i act as correctness forecasts satisfying a weaker-than-oracle (but still unrealistic) notion of conditional calibration. Informally, we require that

$$\mathbb{P}[\boldsymbol{c}_i \text{ is factual} \,|\, \boldsymbol{s}_{\le i}; \text{previous claims } \boldsymbol{c}_{<i} \text{ are all factual}] = s_i. \tag{4}$$

That is, the score is a calibrated forecast of the correctness of the present claim given that all preceding claims were factual. Under this assumption and some additional technical details (see Appendix A for a complete treatment), one can then make the following observation.

Observation 1 (Idealized validity; informal)**.** Under an idealized, stepwise-calibrated[3] scoring function $s(\cdot)$, sampling $y \sim p^{1-\alpha}_{\mathrm{LLM}}(\cdot \,|\, x)$ induced by $\Phi^{1-\alpha}(y \,|\, x)$ satisfies $\mathbb{E}[\ell(x, y)] \le \alpha$.

Intuitively, if the scores were stepwise-calibrated forecasts of conditional correctness given the preceding prefix, a martingale argument (Appendix A) would yield $\prod_{i\le K} s_i = \mathbb{P}(y \text{ correct} \,|\, x, \boldsymbol{s}_{\le K})$. The product of scores then acts as a sequential probability-of-correctness estimate, and the choice $\tau = 1 - \alpha$ directly (conservatively) controls risk at level α.

In practice, heuristic scores are far from calibrated in this sense, and it is generally impossible to achieve calibrated forecasts that are this strong [Van Calster et al., 2016, Hébert-Johnson et al., 2018]. We must therefore choose τ from data via a procedure that makes *essentially no assumptions on the scorer*; this is the main theoretical subject of this paper, which we return to in Section 3.

[2] In this paper, we will routinely use the notation $\boldsymbol{u}_{\le i}$ and $\boldsymbol{u}_{<i}$ to denote prefix vectors of sequences $(u_i)_i$.

[3] By "stepwise-calibrated", we mean that each claim-level score is calibrated as a forecast of the current claim's correctness, conditional on the prompt, the previous scores, and the event that all strictly earlier claims were correct. We also assume that later scores tell us nothing new about whether earlier claims were correct; see Appendix A.

2.3 Calibrated Posterior Sampling on a Mixture Prior

Before we turn to the question of calibrating the threshold τ, we note that our discussion so far relies on the assumption that the posterior is well-defined for all x and τ.

Formally, this amounts to requiring $Z^\tau_{\mathrm{LLM}}(x) = \int_{y \in \mathcal{V}^*} \Phi^\tau(y \mid x)\, \mathrm{d}p_{\mathrm{LLM}}(y \mid x) > 0$ for all $x \in \mathcal{X}$ and $\tau \in [0, 1]$. Since this would require the ability to generate samples that pass the potential's thresholding mechanism by mapping to thresholds that are arbitrarily close to 1, this is too stringent of an assumption on the scoring mechanism. Post-hoc methods such as that of Mohri and Hashimoto [2024] sidestep this issue by treating an all-claims-dropped generation (i.e., outputting the empty set of claims) as an implicit abstention that is always factual. In our setting, p^τ_{LLM} inherits its support from p_{LLM}, which may not include an explicit abstention path. The language model prior p_{LLM} may have no practical amounts of mass on any score-passing outputs at all; strict thresholds thus leave no direct route to safe output, prohibiting calibration. To get around this issue, we therefore calibrate not against p_{LLM} itself but against the *mixture prior*

$$p_\beta(y \mid x) = (1 - \beta)\, p_{\mathrm{LLM}}(y \mid x) + \beta\, \delta_{\texttt{abstain}}(y), \tag{5}$$

where $\delta_{\texttt{abstain}}$ is a point mass on a designated abstention response and $\beta > 0$ is a small pre-specified mass. By convention, we also require that the scorer assigns $S(\texttt{abstain} \mid x) = 1$ and that the abstention is accepted at every threshold; that is, $\Phi^\tau(\texttt{abstain} \mid x) = 1$ for all τ. This captures the intuition that abstaining is always factual, and allows the resulting posterior to retain a factual *escape valve* as τ increases.

Analogously to (2), we then define the mixture posterior $p^\tau_\beta(y \mid x) \propto p_\beta(y \mid x)\Phi^\tau(y \mid x)$. Conveniently for what follows, this also makes the conditional normalizing constant well behaved, i.e., $Z^\tau_\beta(x) = \int_{y \in \mathcal{V}^*} \Phi^\tau(y \mid x)\, \mathrm{d}p_\beta(y \mid x) \geq \beta$ for all τ. We often drop the subscript β and just write $p := p_\beta$ and $p^\tau = p^\tau_\beta$, and similarly $Z^\tau = Z^\tau_\beta$. Note that $p^0 = p^0_\beta = p_\beta$ is simply the prior, since thresholding has no effect at $\tau = 0$ when $\mathcal{S} \subseteq (0, 1]$.

2.4 Instantiating the Scoring Function

While the validity of our method is agnostic to the accuracy of the scorer—we obtain risk-control even under very minimal assumptions—the scorer does have an impact on how stringently the system filters out generations (see Appendix D.5 for an empirical investigation). We follow prior work and elicit scores by asking an external language model to rate its confidence in each claim on a scale from 0 to 1 [Mohri and Hashimoto, 2024]. Empirically, we find that this does not yield continuously-distributed scores, instead concentrating on a small, discrete set. Following Mohri and Hashimoto [2024], in order to calibrate to arbitrary targets of factuality $1 - \alpha$, we therefore add a small random perturbation. For an elicited score $\tilde{s}_i$, we set the final score as $s_i \sim U[\max(0, \tilde{s}_i - \gamma), \min(1, \tilde{s}_i + \gamma)]$, where $\gamma = 0.01$ is a small constant, the exact choice of which has no significant downstream effects, so long as it is on the order of the granularity of the elicited scores[4].

While having flexibility on the choice of scorer is convenient, it also presents a methodological challenge. Since $p^\tau(\cdot \mid x)$ depends on both s and τ, calibration via fixed-sequence testing [Angelopoulos et al., 2025] would require specifying a sensible grid of thresholds τ in advance, and drawing a new calibration set for each (s, τ) pair—a prohibitively expensive process that would need to be repeated whenever s is modified. Even setting computational burden aside, how to pick such a grid is not clear a priori, especially because the spread of the scores is unknown and could be concentrated within some small range, making it difficult to select a sensible discretization *a priori*. Instead, we will aim to develop an *off-policy* calibration routine which only requires samples from the (mixture) prior, eliminating the need for expensive on-policy data collection altogether.

3 Methodology: Calibrating the Threshold τ

We now turn to the central question of the paper: given a family Φ^τ of potential functions and a mixture prior p, how should we pick τ so that posterior sampling controls risk at level α, using *only* samples from the base model? We will assume that we have access to a calibration set of prompts

[4]We find that LLM scorers tend to return scores with at most two decimal points of precision.

$\mathcal{C} = \{x_1, \ldots, x_n\}$ from P_X, and that these prompts are exchangeable. We develop our approach in 3 steps: (1) we derive an idealized calibration rule assuming the normalization constants are known; (2) we construct a tractable empirical approximation of the key quantity being controlled that does not rely on this assumption; (3) we prove that calibrating against this approximation yields finite-sample exact risk-controlling guarantees, which asymptotically transfer to the population posterior as the per-prompt calibration data grows.

(1) Known Normalization Constant. Suppose that we had access to an oracle for the *conditional* normalization constant $Z^\tau(x) = \int_{y \in \mathcal{V}^*} \Phi^\tau(y \mid x)\, \mathrm{d}p(y \mid x)$. If that were the case, then we could estimate the posterior risk using importance reweighting. To see this, note first that from Section 2.1, $\frac{p^\tau(y \mid x)}{p(y \mid x)} = \frac{\Phi^\tau(y \mid x)}{Z^\tau(x)}$. This means that instead of sampling directly from the posterior p^τ, we can i.i.d. sample outputs y_i from the mixture prior $p(\cdot \mid x_i)$, keep only samples whose score exceeds τ, and then re-weight their losses to account for the fact that they were sampled from the prior rather than the posterior. To obtain a finite-sample guarantee, we would then only need to add a conservative correction term $1/(\beta(n+1))$ for the unseen test point and conservatively monotonize the objective by lower bounding the likelihood ratio's denominator. This suggests an *idealized* calibration rule of the following form:

$$\hat{\tau}_{\mathrm{id}} = \inf\left\{\tau \;\middle|\; \frac{1}{n+1}\sum_{i=1}^{n} \frac{\Phi^\tau(y_i \mid x_i)}{Z^{S(y_i \mid x_i)}(x_i)}\ell(x_i, y_i) + \frac{1}{\beta(n+1)} \leq \alpha\right\}. \tag{6}$$

Note that $\tau \mapsto Z^\tau(x_i)$ is a *monotone* non-increasing[5] map, and that since Φ^τ is a thresholding indicator, each term in the sum is an upper bound on $\frac{\Phi^\tau(y_i \mid x_i)}{Z^\tau(x_i)}\ell(x_i, y_i)$. Since the denominator does not itself depend on τ, the entire ratio is therefore monotone in τ. Therefore, the left-hand side of the constraint in (6) inherits this monotonicity, yielding our first guarantee on risk control.

Proposition 3.1. *If $x_1, \ldots, x_n, x_{\mathrm{test}}$ are exchangeable samples from P_X and we have oracle access to $Z^\tau(x_i)$ for all $\tau \in [0,1]$, then computing $\hat{\tau}_{\mathrm{id}}$ according to Equation* (6) *and sampling the test point from the* true *posterior $y_{\mathrm{test}} \sim p^{\hat{\tau}_{\mathrm{id}}}(\cdot \mid x_{\mathrm{test}})$ satisfies $\mathbb{E}[\ell(x_{\mathrm{test}}, y_{\mathrm{test}})] \leq \alpha$, where the randomness is over the draw of the calibration prompts, the calibration responses, and the draw from the test-time posterior.*

The proof is deferred to Appendix B.1.

(2) Unknown $Z^\tau(x)$. In practice, rule (6) cannot be computed since the quantity $Z^\tau(x)$ is intractable.[6] A natural next step is to instead estimate, at each calibration prompt, the conditional posterior risk $\mathbb{E}_{y \sim p^\tau(\cdot \mid x_i)}[\ell(x_i, y)]$ from base-model samples. To form such an estimate, we draw M i.i.d. responses $y_{ij} \sim p_{\mathrm{LLM}}(\cdot \mid x_i)$ (aka *"particles"*) for each $i \in [n]$ and form the empirical (mixture) prior

$$\hat{p}(y \mid x_i, T_i) = (1-\beta)\,\frac{1}{M}\sum_{j \in [M]} \delta_{y_{ij}}(y) + \beta\,\delta_{\texttt{abstain}}(y), \tag{7}$$

where we use the shorthand $T_i = (y_{i1}, \ldots, y_{iM})$ and the abstention atom is attached analytically with mass β rather than sampled. This stratification matches the abstention weight to that of the population mixture prior p_β but, crucially, *ensures that there is deterministic abstention mass, making the plug-in estimator well-defined*. The induced *empirical posterior* is

$$\hat{p}^\tau(y \mid x_i, T_i) = \frac{\Phi^\tau(y \mid x_i)\,\hat{p}(y \mid x_i, T_i)}{\hat{Z}^\tau(x_i, T_i)}, \quad \hat{Z}^\tau(x_i, T_i) = \beta + (1-\beta)\,\frac{1}{M}\sum_{j \in [M]} \Phi^\tau(y_{ij} \mid x_i), \tag{8}$$

which has known normalization constant $\hat{Z}^\tau(x_i, T_i)$. Note that the explicit abstention path implies $\hat{Z}^\tau(x_i, T_i) \geq \beta$, so $\hat{p}^\tau$ is well-defined for all $\tau \in [0,1]$.

(3) Calibrating against the empirical posterior. We now know how to construct an empirical approximation $\hat{p}$ of the prior p, which suggests that we should be able to calibrate the corresponding empirical posterior $\hat{p}^\tau$ with a rule similar to (6). To this end, given T_i, we consider the conditional risk under $\hat{p}^\tau$, namely $\mathbb{E}_{y \sim \hat{p}^\tau(\cdot \mid x_i, T_i)}[\ell(x_i, y)]$, and write the shorthand

$$\hat{H}_i(\tau) = \frac{(1-\beta)\frac{1}{M}\sum_{j \in [M]} \Phi^\tau(y_{ij} \mid x_i)\,\ell(x_i, y_{ij})}{\beta + (1-\beta)\frac{1}{M}\sum_{j \in [M]} \Phi^\tau(y_{ij} \mid x_i)}, \tag{9}$$

[5] This is because as τ increases, the probability mass of completions that pass τ can only decrease.

[6] Indeed, computing it requires summing/integrating over all possible completions y that the LM could generate.

Algorithm 1 Off-policy Conformal Risk Control with Path Monotonization

Require: Calibration prompts $x_1, \dots, x_n$; particle budget M; abstention mass β; target level α

for $i = 1, \dots, n$ **do**
  for $j = 1, \dots, M$ **do**
    Draw $y_{ij} \sim p_{\mathrm{LLM}}(\cdot \mid x_i)$, score it to obtain $\Phi^\tau(y_{ij} \mid x_i)$ for all τ, annotate $\ell(x_i, y_{ij})$
  end for
end for
Form $\hat{H}_i(\tau)$ via (9) and $\bar{H}_i(\tau)$ via (10)
return $\hat{\tau} = \inf\left\{ \tau \middle| \frac{1}{n+1} \sum_{i=1}^{n} \bar{H}_i(\tau) + \frac{1}{n+1} \leq \alpha \right\}$.

because the abstention atom contributes $\ell(x, \texttt{abstain}) = 0$ to the numerator and can be omitted. Note that $\hat{H}_i(\tau)$ is a deterministic function of the calibration data alone, with no further randomness over a sampler: calibration simply inspects the support of $\hat{p}^\tau$, it does not draw from the corresponding empirical distribution.

As a step function of τ[7], $\hat{H}_i$ need not be monotone. Intuitively, raising the threshold may prune a correct particle while leaving incorrect particles, transiently raising the average loss among the accepted particles. We therefore restore monotonicity sample-pathwise by bounding $\hat{H}_i$ with a monotone envelope:

$$\bar{H}_i(\tau) = \sup_{\tau' \geq \tau} \hat{H}_i(\tau'), \tag{10}$$

One risk of this monotonization is that the envelope may lead to very conservative risk estimates, therefore adding slack to the true risk; however, we empirically (see Sec. 4) observe that this is not an issue, since $\hat{H}_i$ is close to monotone for moderate M and $\bar{H}_i$ therefore introduces little slack.

The full procedure is given in Algorithm 1. We can now give the following conformal guarantee, which additionally approaches a guarantee on a sample from the population posterior as M grows (c.f. Corollary 3.4).

Theorem 3.2 (Finite-sample risk control for the empirical posterior)**.** *Let prompts $x_1, \dots, x_n, x_{\mathrm{test}}$ be exchangeable, and assume that conditionally, all calibration and test particles are drawn independently with $y_{ij} \sim p_{\mathrm{LLM}}(\cdot|x_i)$ and $(y_{\mathrm{test}})_j \sim p_{\mathrm{LLM}}(\cdot|x_{\mathrm{test}})$. Let $\hat{\tau}$ be the output of Algorithm 1. If $y_{\mathrm{test}} \sim \hat{p}^{\hat{\tau}}(\cdot|x_{\mathrm{test}}, T_{\mathrm{test}})$, then $\mathbb{E}[\ell(x_{\mathrm{test}}, y_{\mathrm{test}})] \leq \alpha$, where the expectation is over the calibration prompts and particles, test prompt and particles, and the draw from the test-time posterior.*

Given base-model samples, per-claim scores, and factuality annotations on the calibration set, Algorithm 1 returns a threshold $\hat{\tau}$ which controls the risk of the empirical-posterior sampler. The proof (in Sec. B.2) follows a standard conformal-risk-control template: $\bar{H}_i$ is a monotone envelope of a plug-in estimator for the empirical-posterior risk, exchangeability of the calibration tuples bounds an "oracle" threshold $\tau^\star$ at level α, and monotonicity transfers this bound from $\tau^\star$ to $\hat{\tau}$.

From empirical to population posterior. Calibrating against $\hat{p}^{\hat{\tau}}$ raises a natural question: does the guarantee transfer to the population posterior $p^{\hat{\tau}}$? Because $\hat{\tau}$ is data-dependent, pointwise convergence at fixed τ is not enough; we need the convergence to be uniform in τ. Luckily, the threshold-indexed family $\{\Phi^\tau(\cdot \mid x) : \tau \in [0,1]\}$ is a one-dimensional threshold class and hence Glivenko-Cantelli, giving us the following result.

Lemma 3.3 (Uniform asymptotic equivalence of conditional risks)**.** *Assuming continuous scores, fix abstention mass $\beta > 0$, and let $T_M = (y_1, \dots, y_M)$ with $y_j \overset{\text{i.i.d.}}{\sim} p_{\mathrm{LLM}}(\cdot \mid x)$. Then for P_X-a.e. x,*

$$\sup_{\tau \in [0,1]} \left| \hat{\mathbb{E}}_{y \sim \hat{p}^\tau(\cdot \mid x, T_M)}[\ell(x,y)] - \mathbb{E}_{y \sim p^\tau(\cdot \mid x)}[\ell(x,y)] \right| \overset{\text{a.s.}}{\to} 0.$$

Note that this statement holds almost surely over the draw of the calibration particles T_M, as $M \to \infty$, uniformly over all $\tau \in [0,1]$. Combined with Theorem 3.2, this gives us the following corollary.

[7] To see this, note that the only way that τ affects $\hat{H}_i(\tau)$ is through the score indicators $\Phi^\tau(y_{ij}|x_i) = \mathbf{1}\{S(y_{ij}|x_i) > \tau\}$. As such, $\hat{H}_i(\tau)$ only changes when τ crosses one of the sampled scores $S(y_{ij}|x_i)$.

Corollary 3.4. *As the number of calibration particles grows, for* $y_{\text{test}} \sim p^{\hat{\tau}_M}(\cdot \mid x_{\text{test}})$ *we have* $\limsup_{M\to\infty} \mathbb{E}[\ell(x_{\text{test}}, y_{\text{test}})] \leq \alpha$.

In other words, the calibration guarantee transfers asymptotically to any deployed sampler that targets the population posterior. Note, however, that practical deployments of posterior sampling often use approximate, particle-based sampling techniques. It is thus plausible that such approximate samplers will benefit from the finite-sample guarantee (Theorem 3.2) without needing to enter the asymptotic regime, by virtue of being closer to an empirical posterior used for calibration than to the population posterior. We leave investigating this to future work. See Sections 5 and 6 for further discussion on connections between our work and efficient sampling methods such as particle filtering.

4 Experiments

We now empirically evaluate our framework's validity (that is, whether it controls the level of risk at test-time) and utility (whether the resulting generations are useful). We focus on two distinct case studies: one within the hallucination-prone task of open-ended biography generation, and one within mathematical reasoning where inter-claim dependencies are stronger.

Datasets and models. For the biography generation, we use 682 entities from the FActScore dataset [Min et al., 2023]; for mathematical reasoning, we use MATH-500 [Lightman et al., 2024], a standard subset of 500 problem instances from the MATH test set [Hendrycks et al., 2021]. We average over 10 random splits of the data, with 50 instances used as the test set and the remaining used for calibration; results are reported as the mean over the splits, with shaded areas showing approximate 95% CIs using the studentized Wald interval. For generation we use `Llama-3.3-70B-Instruct`[8] [Grattafiori et al., 2024], a model with intermediate capacity on these tasks. We serve the model on a single NVIDIA RTX A6000 with 48GB of VRAM, taking ∼30s per prompt to generate answers. To elicit *scores*, we use OpenAI's GPT-4o, a larger and more capable (but still unreliable) estimator of truth; Appendix D.5 compares against a raw token log-probability scorer and the cheaper GPT-4o-mini. Finally, to ensure that the per-claim *labels* are as accurate as possible, we label the data using GPT-5.4; for FActScore, we include a copy of the entity's Wikipedia article in the prompt.

Base generations and mixture prior setup. We sample 20 base generations from p_{LLM} for each instance, using a moderate temperature of 0.8 (prompts in Appendix E.1). We fix $\beta = 0.05$ for both datasets and report plug-in estimates of the metrics using the empirical posterior over each instance's 20 generations. For example, to compute the empirical risk over a test set $D_{\text{test}} = \{x_i\}_{i=1}^N$, we average (9) over the test instances x_i. We note that for a fixed and finite number of calibration samples M, at very small values of β, the asymptotic guarantees of Corollary 3.4 may not translate to the population posterior p^τ; for a thorough investigation, see Appendix D.8.

Baseline. We compare against Mohri and Hashimoto [2024], the closest prior work. We re-implement Mohri and Hashimoto [2024]'s algorithm so that it does not require the use of a separate model to decompose generations into atomic claims; instead, we encourage the generator to keep clauses brief and split on sentences. We do so to illustrate a tight integration of each method into the generation loop without any reliance on additional restructuring steps, and to enable a direct comparison to our method. For brevity, we refer to this baseline as `MH`. For completeness, we also validate that running `MH` with its original reconstruction step does not change our conclusions (Appendix D.9).

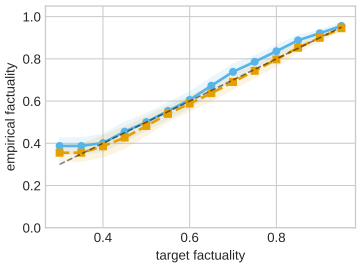
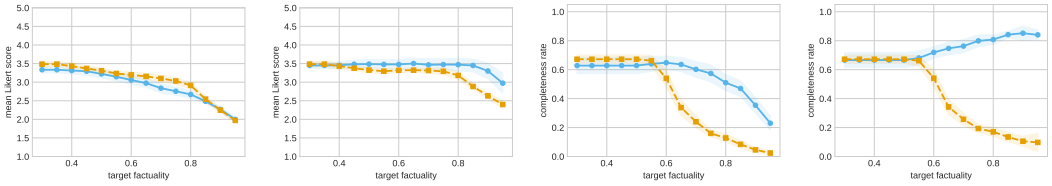


Figure 1. Empirical vs. target factuality on FActScore; ours (—●—) and `MH` (-■-), vs. the ideal relationship $y = x$ (----). Both attain valid coverage.

4.1 Validity

Figure 1 shows the empirical factuality obtained versus the desired target $1-\alpha$ on FActScore (results on MATH in Appendix D.1). As expected from Theorem 3.2, our method achieves valid coverage at all levels. Coverage is also empirically tight; monotonization of the objective does not introduce excessive slack. Note that low targets (≈ 0.4 for FActScore, ≈ 0.5

[8]To reduce cost, we use the (4-bit weights, 16-bit activations)-quantized variant provided by RedHatAI on `HuggingFace`.



(a) FActScore Likert. **(b)** Likert | non-abstain. **(c)** MATH completeness. **(d)** Complete | non-abstain.

Figure 2. LLM-judge utility for ours (—●—) vs. MH (- -■- -) as target factuality varies. Unconditional panels (a, c) include abstentions (judged on FActScore, marked incomplete for MATH); conditional panels (b, d) show results for both methods conditioned on the output not being an abstention.

for MATH) fall below the base model's capacity, so both methods reduce to sampling directly from the prior and therefore flatline. The remaining gap between the two is due to β. As our method samples from the mixture prior $\hat{p}$ instead of p_{LLM}, it is slightly more conservative at targets below the model's base capacity.

4.2 Utility

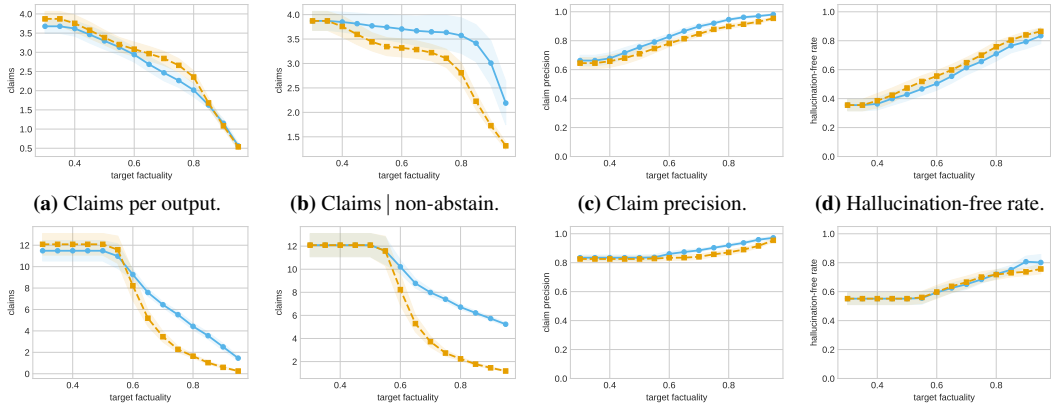
We use LLM-as-a-judge evaluations as our primary measure of the downstream utility of the samples we obtain. For FActScore we construct a Likert-style questionnaire along three axes (*completeness*, *fluency*, and *helpfulness*; rubric in Appendix E.4.1) and ask GPT-5.4 to rate its level of agreement between 1-5 for each. Abstentions are passed to the judge as a pre-formatted response of the form "*I don't know who or what {entity} is.*" For MATH, we simply ask GPT-5.4 whether the proof is *complete*, that is, whether it leads to a final answer and that final answer matches the reference (prompt in Appendix E.4.2). To make sure that our conclusions are not an artifact of the specific choice of judge, Appendices D.3 and D.4 repeat the experiment with Claude Sonnet 5 and, on FActScore, with 200 human raters. All three show broadly consistent patterns in the utility differences between our method and MH. Abstentions are naturally scored as incomplete. To reduce annotation cost, we subsample a single generation[9] for each prompt x_i for this experiment.

In addition to these metrics, we also compute a series of proxy metrics on all the outputs y_{ij}: (i-ii) the average number of claims per output, both with abstentions counting as zero claims and conditioned on non-abstention; (iii) the total fraction of emitted claims that are true, aggregated over all prompts; and (iv) the fraction of non-abstaining outputs that are *entirely* factual, with no false claims at all.

Downstream results. Figure 2 reports the FActScore Likert and MATH completeness scores, both unconditionally (with abstentions scored as described above) and conditioned on obtaining a non-abstaining sample. *FActScore:* Non-abstaining generations from our method degrade less sharply than those from MH as the target factuality increases; the per-axis breakdown in Figure 5 (Appendix D.2) reveals that this is predominantly driven by less degradation in completeness and helpfulness, rather than fluency. In the unconditional average, which conflates response *quality* with response *frequency* (abstention rates are reported in Appendix D.10), our method underperforms MH. This indicates that the judge prefers a highly edited response to refusals in this setting; how such a refusal should be valued relative to a riskier-but-partially-informative answer ultimately depends on the context in which the system is deployed, and this may not reflect the preference of actual users. *MATH:* With a more objective notion of utility, the gains from our method become clearer: while the mixture prior initially increases the probability of sampling an abstaining (and thus incomplete) response, MH begins to generate incomplete answers at very high rates for factuality targets above 0.6. Our method not only *degrades less*, but in the instances where it does not abstain, the rate of completeness even *increases* with the target factuality. These findings thus suggest that our method produces useful generations even at high target levels of factuality. Appendix D.11 shows paired example outputs from both methods.

Proxy results. Figure 3 shows the claim-level proxy metrics, with FActScore on the top row and MATH on the bottom. *First column*: At intermediate targets, our method emits roughly as many claims per instance as MH. The mixture prior places β of the mass on zero-claim abstentions, deflating our method's average at low targets; this reverses at high targets, where MH increasingly produces

[9]For our method, from the posterior; for MH, from the base model with subsequent filtering.



(a) Claims per output. **(b)** Claims | non-abstain. **(c)** Claim precision. **(d)** Hallucination-free rate.

(e) Claims per output. **(f)** Claims | non-abstain. **(g)** Claim precision. **(h)** Hallucination-free rate.

Figure 3. Claim-level proxy metrics for ours (—●—) vs. MH (- -■- -) on FActScore (top) and MATH (bottom).

heavily filtered generations. *Second column:* Restricting our attention to non-abstaining responses, our method emits substantially more claims on both datasets. Intuitively, MH distributes rejected claims evenly across instances, while our method concentrates filtering on hard instances and leaves easier prompts with full-length generations. *Third column*: The fraction of emitted claims which are correct is similar for both methods, increasing (as expected) with the target level. *Fourth column*: The one axis on which our method may lose ground is the fraction of non-abstaining outputs that are *entirely* factual. On FActScore in particular, we observe a slight decrease in the chance that *conditioned on not being an abstention*, any given sample is completely factual. Ultimately, whether non-abstaining, hallucination-free responses can be obtained is controlled by the capacity of the underlying p_{LLM}: posterior sampling can only shift the mass of the samples towards safer regions if such regions of support exist.

Summary. While both our method and MH control similar objectives, the resulting downstream utility differs significantly. As the target level of risk control increases, our method produces more helpful and complete outputs (as judged by an external LLM).[10]

However, since our sampler cannot stray far from the support of p_{LLM}, instances poorly known to the LLM can only be pushed towards factuality by abstaining, leading to a small uptick in the chance that any given *non-abstaining* answer has at least one incorrect claim in settings where the model lacks the capacity to solve the task.

5 Related Work

Calibrated stopping rules. A complementary body of recent work calibrates stopping rules for reasoning models and agents [Wu et al., 2025a, Sadhuka et al., 2025, Zhou et al., 2026]. These methods similarly avoid the limitations of post-hoc filtering (Section 1), but the target of calibration is different: they calibrate decision rules over the reasoning process (when to stop thinking, or when to reject an agent trajectory), whereas we calibrate a family of *output distributions*. On the methodological side, these methods respectively build upon fixed sequence testing [Angelopoulos et al., 2025] and E-values [Ramdas and Wang, 2025], and leverage on-policy samples; this is impractical in our setup, where the distribution of scores can be arbitrarily bimodal and no natural grid is available in advance. The goals and approaches of these works are thus orthogonal to ours.

Test-time scaling and unreliable process rewards. Viewing the score function as an unreliable *process reward* [Uesato et al., 2022, Lightman et al., 2024, Wang et al., 2024], our method can be seen as a form of test-time scaling [Snell et al., 2025, Brown et al., 2024, Wu et al., 2025b]. Our approach of calibrating a truncated distribution is also reminiscent of a line of work that frames semantically constrained generation as posterior sampling [Lew et al., 2023, Zhao et al., 2024, Loula et al., 2025]. One salient difference between such works and ours is that unlike typical test-time scaling or control, in which one wants to efficiently sample from a *fixed* (but possibly unknown) reward-tilted distribution, in our setting we seek to *calibrate* the distribution to obtain a desired level of risk

[10] See Appendix D.7 for an investigation of how the choice of score function affects the extent to which utility gains arise from resampling vs. from filtering at the level of whole outputs instead of individual claims.

control. However, these goals often go hand in hand, and preliminary experiments (see Appendix D.7) suggest that increasing M—the parallel test-time budget—can increase downstream utility as well as maintaining risk control. Our work is thus complementary to efforts combining unreliable process rewards with efficient approximate inference techniques such as particle filtering [Puri et al., 2025, Golowich et al., 2026]. Combining risk control with such efficient samplers is left for future work.

Conformal prediction and risk control under distribution shift. Conformal methods [Saunders et al., 1999, Angelopoulos et al., 2024a] have widely been adapted to work under known distribution shift [Tibshirani et al., 2019, Angelopoulos et al., 2024b], tackling the issue via weighted exchangeability. While we draw inspiration from this approach, we face an additional complication: label distribution shift that is itself directly affected by the calibration parameter. This prevents us from applying methods based on vanilla variants of weighted exchangeability; see Appendix C, as well as the works of Fannjiang et al. [2022] and Prinster et al. [2024] for a more thorough treatment of conformal guarantees under general (covariate) shifts. In contemporary work, Prinster et al. [2026] proposed *Conformal Policy Control* (CPC), which shares many conceptual and motivational aspects with ours. When applied to language modeling, CPC can be used to interpolate between a safe policy and a high-utility model, with the strength of interpolation set via a modified conformal risk control procedure. While this achieves a very similar goal as our method, it induces a different safety-utility tradeoff since we tilt the distribution towards high conformal scores rather than towards a separate language model policy.

Extensions of Conformal Prediction. Finally, we mention that there are many extensions to conformal prediction, such as approximate-conditional variants [Gibbs et al., 2025] and ones robust to adversarial distribution shifts [Gibbs and Candes, 2021]. We believe that our method could accommodate such extensions; as a preliminary experiment, Appendix D.6 investigates the group-conditional setting, finding that it outperforms prior work based on filtering Cherian et al. [2024].

6 Discussion: Limitations & Future Directions

The β parameter. This parameter governs the prior mixture between the model's own responses and the explicit abstention response. Its main role is to ensure that the posterior remains well-defined by assigning fixed mass to an always-factual abstention atom, guaranteeing that the normalization constant is bounded away from zero. As a result, β controls a *utility-abstention tradeoff*, but not the validity of the procedure on the empirical posterior $\hat{p}^{\hat{\tau}}$. Larger β values provide a stronger fallback path, but shift more probability mass toward abstention and reduce the likelihood of sampling a non-abstaining output. This primarily has an effect on the convergence rate of the empirical posterior $\hat{p}^{\hat{\tau}}$ to the population posterior $p^{\hat{\tau}}$ (c.f. Corollary 3.4), which is affected by the scale of β compared to M. If convergence to the population posterior is not desired, we recommend setting a small β, and only abstaining when no sample passing the scoring threshold was found.

Measuring downstream utility. Due to cost, we mainly evaluate downstream utility by asking an external LLM to judge responses on properties such as completeness, fluency and mathematical coherence. We complement these experiments with a small human study (Appendix D.3). While this provides a scalable first validation of our method, it does not completely guarantee that our measure of downstream utility agrees with users' opinions of each system. In particular, a fully human study with a more thorough rubric could shed more light on the trade-off between the utility of responses vs abstentions, as discussed in Section 4.

Efficient posterior sampling. Another practical consideration is how posterior particles are sampled. In our experiments, the posterior is approximated using finite sets of base-model completions sampled from the underlying model, together with the abstention atom. These particles are rescored and reweighted to form an empirical posterior, which is sufficient to validate the finite-sample guarantees proved in the paper. However, this approximation may become inefficient when high-scoring completions are rare or concentrated in low-probability regions of the base distribution. Developing *more efficient* posterior samplers is therefore an important direction for future work.

Acknowledgments and Disclosure of Funding

The authors thank Stephen Bates, Vivek Farias, Adam Fisch and Drew Prinster for insightful discussions. We would also like to thank the reviewers for helpful feedback and additional suggestions. NE and CP were partially supported by an MIT-Google grant. ASL and TXO were supported by the National Science Foundation under Grant No. 1918839.

References

Anastasios Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International conference on learning representations*, volume 2024, pages 55198–55218, 2024a.

Anastasios N Angelopoulos, Rina Foygel Barber, and Stephen Bates. Theoretical foundations of conformal prediction. *arXiv preprint arXiv:2411.11824*, 2024b.

Anastasios N Angelopoulos, Stephen Bates, Emmanuel J Candès, Michael I Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19(2):1641–1662, 2025.

Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V. Le, Christopher Ré, and Azalia Mirhoseini. Large language monkeys: Scaling inference compute with repeated sampling. *arXiv preprint arXiv:2407.21787*, 2024.

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

Mehul Damani, Isha Puri, Stewart Slocum, Idan Shenfeld, Leshem Choshen, Yoon Kim, and Jacob Andreas. Beyond binary rewards: Training LMs to reason about their uncertainty. In *International Conference on Learning Representations*, 2026.

Fatemeh Dehghani, Roya Dehghani, Yazdan Naderzadeh Ardebili, and Shahryar Rahnamayan. Large language models in legal systems: A survey. *Humanities and Social Sciences Communications*, 12 (1):1977, 2025.

Gianluca Detommaso, Martin Bertran, Riccardo Fogliato, and Aaron Roth. Multicalibration for confidence scoring in LLMs. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Dongyang Fan, Sebastien Delsad, Nicolas Flammarion, and Maksym Andriushchenko. HalluHard: A hard multi-turn hallucination benchmark. *arXiv preprint arXiv:2602.01031*, 2026.

Clara Fannjiang, Stephen Bates, Anastasios N Angelopoulos, Jennifer Listgarten, and Michael I Jordan. Conformal prediction under feedback covariate shift for biomolecular design. *Proceedings of the National Academy of Sciences*, 119(43):e2204569119, 2022.

Farieda Gaber, Maqsood Shaik, Fabio Allega, Agnes Julia Bilecz, Felix Busch, Kelsey Goon, Vedran Franke, and Altuna Akalin. Evaluating large language model workflows in clinical decision support for triage and referral and diagnosis. *npj Digital Medicine*, 8(1):263, 2025.

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

Isaac Gibbs, John J Cherian, and Emmanuel J Candès. Conformal prediction with conditional guarantees. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87(4): 1100–1126, 2025.

Noah Golowich, Fan Chen, Dhruv Rohatgi, Raghav Singhal, Carles Domingo-Enrich, Dylan J Foster, and Akshay Krishnamurthy. Reject, resample, repeat: Understanding parallel reasoning in language model inference. *arXiv preprint arXiv:2603.07887*, 2026.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Ursula Hébert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning*. PMLR, 2018.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021.

Alexander K Lew, Tan Zhi-Xuan, Gabriel Grand, and Vikash K Mansinghka. Sequential Monte Carlo steering of large language models using probabilistic programs. *arXiv preprint arXiv:2306.03081*, 2023.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=v8L0pN6EOi`.

João Loula, Benjamin LeBrun, Li Du, Ben Lipkin, Clemente Pasti, Gabriel Grand, Tianyu Liu, Yahya Emara, Marjorie Freedman, Jason Eisner, et al. Syntactic and semantic control of large language models via sequential Monte Carlo. In *International Conference on Learning Representations*, volume 2025, pages 64758–64791, 2025.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 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*, pages 12076–12100, 2023.

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

Drew Prinster, Samuel Don Stanton, Anqi Liu, and Suchi Saria. Conformal validity guarantees exist for any data distribution (and how to find them). In *Proceedings of the 41st International Conference on Machine Learning*, pages 41086–41118, 2024.

Drew Prinster, Clara Fannjiang, Ji Won Park, Kyunghyun Cho, Anqi Liu, Suchi Saria, and Samuel Don Stanton. Conformal policy control. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Isha Puri, Shivchander Sudalairaj, Guangxuan Xu, Abhishek Bhandwaldar, Kai Xu, and Akash Srivastava. Rollout roulette: A probabilistic inference approach to inference-time scaling of LLMs using particle-based Monte Carlo methods. *Advances in Neural Information Processing Systems*, 38:156427–156454, 2025.

Victor Quach, Adam Fisch, Tal Schuster, Adam Yala, Jae Ho Sohn, Tommi Jaakkola, and Regina Barzilay. Conformal language modeling. In *International Conference on Learning Representations*, volume 2024, pages 11654–11681, 2024.

Aaditya Ramdas and Ruodu Wang. Hypothesis testing with e-values. *Foundations and Trends® in Statistics*, 1(1-2):1–390, 2025.

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

Shuvom Sadhuka, Drew Prinster, Clara Fannjiang, Gabriele Scalia, Bonnie Berger, Aviv Regev, and Hanchen Wang. E-valuator: Reliable agent verifiers with sequential hypothesis testing. *arXiv preprint arXiv:2512.03109*, 2025.

Craig Saunders, Alexander Gammerman, and Volodya Vovk. Transduction with confidence and credibility. In *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*, 1999.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *International Conference on Learning Representations (ICLR)*, 2025.

Ryan J Tibshirani, Rina Foygel Barber, Emmanuel J Candès, and Aaditya Ramdas. Conformal prediction under covariate shift. *Advances in neural information processing systems*, 32, 2019.

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. Solving math word problems with process- and outcome-based feedback. *arXiv preprint arXiv:2211.14275*, 2022.

Ben Van Calster, Daan Nieboer, Yvonne Vergouwe, Bavo De Cock, Michael J Pencina, and Ewout W Steyerberg. A calibration hierarchy for risk models was defined: From utopia to empirical data. *Journal of clinical epidemiology*, 74:167–176, 2016.

Aad W van der Vaart and Jon A Wellner. *Weak convergence and empirical processes: With applications to statistics*. Springer Series in Statistics. Springer, 1996.

Vladimir Vovk, David Lindsay, Ilia Nouretdinov, and Alex Gammerman. Mondrian confidence machine. Working paper, Royal Holloway University of London, 2003.

Peiyi Wang, Lei Li, Zhihong Shao, R. X. Xu, Damai Dai, Yifei Li, Deli Chen, Y. Wu, and Zhifang Sui. Math-Shepherd: Verify and reinforce LLMs step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 9426–9439, 2024.

Menghua Wu, Cai Zhou, Stephen Bates, and Tommi Jaakkola. Thought calibration: Efficient and confident test-time scaling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025a.

Yangzhen Wu, Zhiqing Sun, Shanda Li, Sean Welleck, and Yiming Yang. Inference scaling laws: An empirical analysis of compute-optimal inference for problem-solving with language models. In *International Conference on Learning Representations (ICLR)*, 2025b.

Stephen Zhao, Rob Brekelmans, Alireza Makhzani, and Roger Baker Grosse. Probabilistic inference in language models via twisted sequential Monte Carlo. In *International Conference on Machine Learning*, pages 60704–60748. PMLR, 2024.

Cai Zhou, Zekai Wang, Menghua Wu, Qianyu Julie Zhu, Flora C Shi, Chenyu Wang, Ashia Wilson, Tommi Jaakkola, and Stephen Bates. Online reasoning calibration: Test-time training enables generalizable conformal LLM reasoning. *arXiv preprint arXiv:2604.01170*, 2026.

A Consequences of Calibrated Scores

In this section, we make the informal statement about calibrated forecasts from Section 2.2 precise. Define the random variable $Z_i := \mathbf{1}\{\text{claim } i \text{ is conditionally correct}\}$, and the indicator $E_i := \mathbf{1}\{\text{claim } i \text{ ends with an } \texttt{EOS} \text{ token}\}$. We will reason about the filtration $\mathbb{F} = \{\mathcal{F}_i\}_{i\in\mathbb{N}}$ where $\mathcal{F}_i := \sigma(x, \boldsymbol{s}_{\leq i}, \mathbf{E}_{\leq i})$. Note that the inclusion of information about end of sequence tokens will allow us to make statements about random stopping times (in particular the random K) and can be thought of as a technical artefact void of any deeper meaning.

For shorthand, let $Z_{<i} = \prod_{j=1}^{i-1} Z_j$ and $Z_{\leq i} = Z_{<i} Z_i$. Suppose then that the scores were such that

$$\mathbb{E}[Z_i \mid \mathcal{F}_i, Z_{<i} = 1] = s_i \qquad \text{for } i \geq 1. \tag{11}$$

Intuitively, this means that conditioned on all prior scores and the prior claims being correct, the score s_i is a calibrated predictor of the correctness of the subsequent claim c_i. When $i > K$, it is trivially satisfied.

Suppose also that the current score s_i does not give any additional information about the correctness of *past* claims that is not already captured by the preceding scores $\boldsymbol{s}_{<i}$; that is,

$$\mathbf{1}\{Z_{<i} = 1\} \perp s_i, E_i \mid \mathcal{F}_{i-1}. \tag{12}$$

Intuitively, this means that the correctness of a *prefix* can be asserted without looking at the rest of the sentence. We will also write $\mathbf{1}\{Z_{<i} = 1\} \perp \mathcal{F}_i \mid \mathcal{F}_{i-1}$. This essentially assumes that future scores give no additional information about past correctness, if the past scores are available, which we find plausible[11]. Note also that if $\mathcal{F}_{k-1}$ reveals that some claim $c_1, \ldots, c_{k-1}$ contains an $\texttt{EOS}$ token, then we assume for mathematical convenience that s_k is deterministically 1 conditional on the information in $\mathcal{F}_{k-1}$. Then, no information is gained from $\mathcal{F}_k$ either, and the assumption is also trivially satisfied when $k > K$.

This appendix supplies the formal statement and derivation backing the informal Observation 1 in Section 2.2. Concretely:

Proposition A.1 (Formal version of Observation 1)**.** *Let $\{s_i\}_{i\geq 1}$ be the claim-level score sequence produced when sampling $y \sim p_{\mathrm{LLM}}(\cdot \mid x)$, let K denote the random number of claims in y, and let Z_i indicate conditional correctness of claim i (Section 2.2). Suppose s satisfies the stepwise-calibration condition* (11) *and the conditional independence condition* (12)*. Then for any $\tau \in [0, 1]$,*

$$\mathbb{P}\left(Z_{\leq K} = 1 \;\middle|\; \prod_{i=1}^{K} s_i > \tau\right) \geq \tau.$$

Setting $\tau = 1 - \alpha$ gives $\mathbb{E}[\ell(x, y) \mid \Phi^{1-\alpha}(y \mid x) = 1] \leq \alpha$.

The remainder of the appendix develops the argument. We pick up the discussion from Section 2.2 again. Recall that we defined the σ-algebras $\mathcal{F}_i := \sigma(x, \boldsymbol{s}_{\leq i}, \mathbf{E}_{\leq i})$. Formally, we define the stopping time K to be

$$K = \inf\{k \mid E_k = 1\}.$$

Making what we mentioned above about mathematical convenience more precise, we also set $Z_k = 1$, $s_k = 1$ and $c_k = \emptyset$ for $k > K$. Clearly, K is a valid stopping time, since the event $\{K \leq \ell\}$ is $\mathcal{F}_\ell$-measurable, because $\mathcal{F}_\ell$ includes $\mathbf{E}_{\leq \ell}$ in its definition, and the information contained in the events $E_1, \ldots, E_\ell$ is sufficient to determine whether $K = \ell$ has occurred.

Let us first fix k.

[11] And in any case, this development is purely motivating, and not required for the correctness and calibration guarantees in Section 3.

$$\begin{aligned}
\mathbb{E}[Z_{\leq k} \mid \mathcal{F}_k] &= \mathbb{E}[Z_{<k} Z_k \mid \mathcal{F}_k] \\
&= \mathbb{E}[Z_{<k} \mathbb{E}[Z_k \mid \mathcal{F}_k, Z_{<k}] \mid \mathcal{F}_k] \\
&\overset{(i)}{=} \mathbb{E}[Z_{<k} \mathbb{E}[Z_k \mid \mathcal{F}_k, Z_{<k} = 1] \mid \mathcal{F}_k] \\
&\overset{(ii)}{=} \mathbb{E}[Z_{<k} s_k \mid \mathcal{F}_k] \\
&= s_k \, \mathbb{P}(Z_{<k} = 1 \mid \mathcal{F}_k) \\
&\overset{(iii)}{=} s_k \, \mathbb{P}(Z_{<k} = 1 \mid \mathcal{F}_{k-1}).
\end{aligned}$$

We briefly comment on the steps:

(i) follows because

$$\mathbb{E}[Z_k \,|\, \mathcal{F}_k, Z_{<k}] = \mathbf{1}\{A\} \, \mathbb{E}[Z_k \,|\, \mathcal{F}_k, A] + \mathbf{1}\{A^c\} \, \mathbb{E}[Z_k \,|\, \mathcal{F}_k, A^c]$$

and for any event A that is determined by $Z_{<k}$ (meaning deterministic conditional on the realization of the random variable $Z_{<k}$). Multiplying by $\mathbf{1}\{A\}$ for $A = \{Z_{<k} = 1\}$ shows the step.

(ii) follows because of (11)

(iii) follows because of (12)

A proof by induction shows that for any *fixed* $k \geq 1$

$$\mathbb{E}[Z_{\leq k} \mid \mathcal{F}_k] = \textstyle\prod_{i=1}^{k} s_i. \tag{13}$$

Since K is a stopping time and $K \leq L$ almost surely[12], we can deduce that for the stopped σ-algebra,

$$\mathbb{E}[Z_{\leq K} \mid \mathcal{F}_K] = \textstyle\prod_{i=1}^{K} s_i, \tag{14}$$

which shows the first part of our claim.

We are now interested in

$$\mathbb{E}\left[\ell(x,y) \,|\, \prod_{i=1}^{K} s_i > \tau\right] = \mathbb{P}\left(\ell(x,y) = 1 \,|\, \prod_{i=1}^{K} s_i > \tau\right) = \mathbb{P}\left(Z_{\leq K} = 0 \,|\, \prod_{i=1}^{K} s_i > \tau\right).$$

In words, this is the probability of incorrectness of some claim jointly over the claims, when we sample from the posterior distribution conditional on the scores being above the threshold τ. We will now show that this quantity is in turn upper bounded by $1 - \tau$, suggesting that calibrated scores lead to the choice $\tau = 1 - \alpha$ and guarantee α-level risk control.

To see why, let us first rewrite

$$\begin{aligned}
\mathbb{P}(Z_{\leq K} = 1 \,|\, \textstyle\prod_{i=1}^{K} s_i > \tau) &= \frac{\mathbb{P}(Z_{\leq K} = 1 \text{ and } \prod_{i=1}^{K} s_i > \tau)}{\mathbb{P}(\prod_{i=1}^{K} s_i > \tau)} \\
&= \frac{\mathbb{E}\left[\mathbf{1}\left\{Z_{\leq K} = 1 \text{ and } \prod_{i=1}^{K} s_i > \tau\right\}\right]}{\mathbb{P}(\prod_{i=1}^{K} s_i > \tau)} \\
&= \frac{\mathbb{E}\left[Z_{\leq K} \mathbf{1}\left\{\prod_{i=1}^{K} s_i > \tau\right\}\right]}{\mathbb{P}(\prod_{i=1}^{K} s_i > \tau)}.
\end{aligned}$$

We focus on the numerator. Since $\prod_{i=1}^{K} s_i$ is $\mathcal{F}_K$-measurable for the random stopping time K, the event $A_\tau = \left\{\prod_{i=1}^{K} s_i > \tau\right\}$ is also $\mathcal{F}_K$-measurable. Therefore

$$\begin{aligned}
\mathbb{E}[Z_{\leq K} \mathbf{1}_{A_\tau}] &= \mathbb{E}[\mathbb{E}[Z_{\leq K} \mathbf{1}_{A_\tau} \mid \mathcal{F}_K]] \\
&= \mathbb{E}[\mathbb{E}[Z_{\leq K} \mid \mathcal{F}_K] \mathbf{1}_{A_\tau}] \\
&= \mathbb{E}[\textstyle\prod_{i=1}^{K} s_i \, \mathbf{1}_{A_\tau}] \\
&\geq \tau \mathbb{E}[\mathbf{1}_{A_\tau}] \\
&= \tau \mathbb{P}(A_\tau).
\end{aligned}$$

[12]The maximum length of the LLM response

Dividing by $\mathbb{P}(A_\tau)$ gives

$$\mathbb{E}[Z_{\leq K} \mid A_\tau] \geq \tau.$$

Since $Z_{\leq K}$ is binary, and setting $\tau = 1 - \alpha$, this is equivalent to

$$\mathbb{P}(Z_{\leq K} = 1 \mid \Phi^{1-\alpha}(y \mid x) = 1) \geq 1 - \alpha.$$

B Deferred Proofs from Section 3

B.1 Proof of Proposition 3.1 (Known $Z^\tau(x)$)

We collect the auxiliary results needed for the proof, then present the main argument. Along with monotonicity by design, the proof will make use of the following likelihood ratio reweighting:

Fact 1. Fix x and τ. Suppose we draw $y \sim p(\cdot \mid x)$ and $y' \sim p^\tau(\cdot \mid x)$. Then

$$\mathbb{E}_{y \sim p(\cdot \mid x)} \left[\frac{\Phi^\tau(y \mid x)}{Z^\tau(x)} \ell(x, y) \right] = \mathbb{E}_{y' \sim p^\tau(\cdot \mid x)} \left[\ell(x, y') \right]. \tag{15}$$

Proof of Fact 1

Proof. We have

$$\begin{aligned}
\mathbb{E}_{y \sim p^\tau(\cdot \mid x)} \left[\ell(x, y) \right] &= \sum_{y \in \mathcal{V}^*} \ell(x, y)\, p^\tau(y \mid x) \\
&= \sum_{y \in \mathcal{V}^*} \ell(x, y) \frac{\Phi^\tau(y \mid x)\, p(y \mid x)}{Z^\tau(x)} \\
&= \mathbb{E}_{y \sim p(\cdot \mid x)} \left[\ell(x, y) \frac{\Phi^\tau(y \mid x)}{Z^\tau(x)} \right].
\end{aligned}$$

□

We now turn to the monotonized importance weights $\bar{w}_i(\tau) = \frac{\Phi^\tau(y_i \mid x_i)}{Z^{S(y_i \mid x_i)}(x_i)}$, as well as their relationship to the weights $w_i(\tau) = \frac{\Phi^\tau(y_i \mid x_i)}{Z^\tau(x_i)}$. We write $\ell_i = \ell(x_i, y_i)$ as a shorthand in what follows.

Lemma B.1. *The monotonized weights $\bar{w}_i(\tau)$ satisfy:*

(1) $\bar{w}_i(\tau)\ell_i$ is non-increasing in τ.

(2) $\bar{w}_i(\tau)\ell_i \geq w_i(\tau)\ell_i$ for all τ, and hence $\mathbb{E}_{p^0}[\bar{w}_i(\tau)\ell_i] \geq \mathbb{E}_{p^\tau}[\ell_i]$.

(3) $\bar{w}_i(\tau)\ell_i \leq 1/\beta$.

(4) $\bar{w}_i(\tau)\ell_i$ is right-continuous in τ.

Proof. (1): $\bar{w}_i(\tau)\ell_i = \frac{\ell_i}{Z^{S(y_i \mid x_i)}(x_i)} \mathbf{1}\{S(y_i \mid x_i) > \tau\}$ is constant for $\tau < S(y_i \mid x_i)$ and zero for $\tau \geq S(y_i \mid x_i)$. (2): For $\tau < S(y_i \mid x_i)$, $Z^{S(y_i \mid x_i)}(x_i) \leq Z^\tau(x_i)$ (since Z^τ is non-increasing), so $\bar{w}_i(\tau) \geq w_i(\tau)$; for $\tau \geq S(y_i \mid x_i)$, both are zero. The expectation bound follows from Fact 1. (3): $Z^{S(y_i \mid x_i)}(x_i) \geq \beta$ since the abstention atom always survives, and $\ell, \Phi^\tau \in \{0, 1\}$. (4): This follows immediately from the same reasoning as (1), since the expression is constant on $[\tau, S(y_i \mid x_i))$ for $\tau < S(y_i \mid x_i)$ and on $[\tau, 1]$ otherwise. □

The proof of Proposition 3.1 assumes a hallucinated completion from the mixture prior on the test prompt. Formally, we assume that $y_{n+1} \overset{\text{i.i.d.}}{\sim} p(\cdot \mid x_{\text{test}})$ conditionally independently of the sample y_{test} given x_{test}. We also denote $x_{n+1} = x_{\text{test}}$ for convenience. The selection rule $\hat{\tau}_{\text{id}}$ is then compared to the "oracle" one that assumes access to a hallucinated base completion:

$$\tau^*_{\text{id}} = \inf \left\{ \tau \mid \frac{1}{n+1} \sum_{i=1}^{n+1} \frac{\Phi^\tau(y_i \mid x_i)}{Z^{S(y_i \mid x_i)}(x_i)} \ell(x_i, y_i) \leq \alpha \right\}. \tag{16}$$

Note that this definition is well-posed. The set on the right-hand side always contains $\tau = 1$ (only the abstention passes the threshold there, and it incurs no loss), so the feasible set is non-empty. Furthermore, by Lemma B.1 (1) and (4) its defining function is non-increasing and right-continuous in τ, so the infimum both exists and is attained.

Lemma B.2. $\hat{\tau}_{\mathrm{id}} \geq \tau^*_{\mathrm{id}}$ *almost surely.*

Proof. For any τ in the feasible set of $\hat{\tau}_{\mathrm{id}}$, we have $\frac{1}{n+1}\sum_{i=1}^{n} \bar{w}_i(\tau)\ell_i + \frac{1}{\beta(n+1)} \leq \alpha$. Since $\bar{w}_{n+1}(\tau)\ell_{n+1} \leq 1/\beta$ by Lemma B.1:

$$\frac{1}{n+1}\sum_{i=1}^{n+1} \bar{w}_i(\tau)\ell_i \leq \frac{1}{n+1}\sum_{i=1}^{n} \bar{w}_i(\tau)\ell_i + \frac{1}{\beta(n+1)} \leq \alpha.$$

So τ is in the feasible set of τ^*_{id}, giving $\tau^*_{\mathrm{id}} \leq \hat{\tau}_{\mathrm{id}}$ a.s. □

Main Proof of Proposition 3.1

Proof. Step 1: Reduce to monotonized weights at $\hat{\tau}_{\mathrm{id}}$. Recall that the test response y_{test} is drawn from $p^{\hat{\tau}_{\mathrm{id}}}(\cdot \mid x_{\mathrm{test}})$ for a fresh test prompt $x_{\mathrm{test}} \sim P_X$ exchangeable with the calibration prompts, where $\hat{\tau}_{\mathrm{id}} = \hat{\tau}_{\mathrm{id}}(D_{1:n})$ depends only on the calibration set. Let us denote by $x_{n+1} = x_{\mathrm{test}}$ and let $y_{n+1} \sim p(\cdot \mid x_{n+1})$ be a *hallucinated* sample from the mixture prior. Since $\hat{\tau}_{\mathrm{id}}$ is chosen without knowledge of the realization of (x_{n+1}, y_{n+1}), Fact 1 gives $\mathbb{E}[\ell(x_{\mathrm{test}}, y_{\mathrm{test}}) \mid \hat{\tau}_{\mathrm{id}}, x_{\mathrm{test}}] = \mathbb{E}[w_{n+1}(\hat{\tau}_{\mathrm{id}})\ell(x_{n+1}, y_{n+1}) \mid \hat{\tau}_{\mathrm{id}}, x_{\mathrm{test}}]$. By Lemma B.1, $\bar{w}_{n+1}(\tau) \geq w_{n+1}(\tau)$ pointwise, so

$$\mathbb{E}[\ell(x_{\mathrm{test}}, y_{\mathrm{test}}) \mid \hat{\tau}_{\mathrm{id}}] \leq \mathbb{E}[\bar{w}_{n+1}(\hat{\tau}_{\mathrm{id}})\, \ell_{n+1} \mid \hat{\tau}_{\mathrm{id}}], \quad \text{a.s.}$$

From this

$$\mathbb{E}[\ell(x_{\mathrm{test}}, y_{\mathrm{test}})] \leq \mathbb{E}[\bar{w}_{n+1}(\hat{\tau}_{\mathrm{id}})\, \ell_{n+1}] \tag{17}$$

where we took the expectation over the randomness in $\hat{\tau}_{\mathrm{id}}$ again.

Step 2: Transfer from $\hat{\tau}_{\mathrm{id}}$ to τ^*_{id}. By Lemma B.2, $\hat{\tau}_{\mathrm{id}} \geq \tau^*_{\mathrm{id}}$ a.s. Since by Lemma B.1 $\bar{w}_{n+1}(\tau)\ell_{n+1}$ is non-increasing in τ:

$$\bar{w}_{n+1}(\hat{\tau}_{\mathrm{id}})\, \ell_{n+1} \leq \bar{w}_{n+1}(\tau^*_{\mathrm{id}})\, \ell_{n+1}, \quad \text{a.s.} \tag{18}$$

Step 3: Exchangeability. The hallucinated calibration set $D_{1:n+1} = \{(x_i, y_i)\}_{i=1}^{n+1}$ is exchangeable[13], and $\tau^*_{\mathrm{id}} = \tau^*_{\mathrm{id}}(D_{1:n+1})$ is permutation invariant. Therefore, conditional on the multiset $E = \{(x_i, y_i)\}_{i \in [n+1]}$, we have

$$\mathbb{E}[\bar{w}_{n+1}(\tau^*_{\mathrm{id}})\, \ell_{n+1} \mid E] = \frac{1}{n+1}\sum_{i=1}^{n+1} \bar{w}_i(\tau^*_{\mathrm{id}})\ell_i \leq \alpha,$$

where inequality uses the definition of τ^*_{id} in (16). Integrating out E again gives us

$$\mathbb{E}[\bar{w}_{n+1}(\tau^*_{\mathrm{id}})\, \ell_{n+1}] = \mathbb{E}\left[\frac{1}{n+1}\sum_{i=1}^{n+1} \bar{w}_i(\tau^*_{\mathrm{id}})\, \ell_i\right] \leq \alpha.$$

Finally, we combine the pieces to obtain

$$\mathbb{E}[\ell(x_{\mathrm{test}}, y_{\mathrm{test}})] \overset{(17)}{\leq} \mathbb{E}[\bar{w}_{n+1}(\hat{\tau}_{\mathrm{id}})\, \ell_{n+1}] \overset{(18)}{\leq} \mathbb{E}[\bar{w}_{n+1}(\tau^*_{\mathrm{id}})\, \ell_{n+1}] \leq \alpha.$$

□

[13] Indeed, we assumed the prompts were exchangeable, and the responses were conditionally independent draws from $p(\cdot \mid x_i)$.

B.2 Proof of Theorem 3.2 (Empirical Posterior Calibration)

We first give an equivalent of Fact 1 for this setting.

Lemma B.3 (Risk under Empirical Posterior)**.** *Let γ be any $[0,1]$-valued random variable. For any i, suppose that conditional on observing (x_i, T_i, γ) we draw $y \sim \hat{p}^\gamma(\cdot \mid x_i, T_i)$. Then*

$$\mathbb{E}[\ell(x_i, y) \mid x_i, T_i, \gamma] = \hat{H}_i(\gamma).$$

Proof. Conditional on (x_i, T_i, γ), the distribution $\hat{p}^\gamma(\cdot \mid x_i, T_i)$ is finitely supported on $\{y_{i1}, \ldots, y_{iM}, \texttt{abstain}\}$ with mass $\frac{(1-\beta)\Phi^\gamma(y_{ij} \mid x_i)/M}{\hat{Z}^\gamma(x_i, T_i)}$ on each y_{ij} and mass $\frac{\beta}{\hat{Z}^\gamma(x_i, T_i)}$ on $\texttt{abstain}$. Summing ℓ against these weights and using $\ell(x_i, \texttt{abstain}) = 0$ yields the result. □

Lemma B.4. *The monotonized estimator $\bar{H}_i(\tau) = \sup_{\tau' \geq \tau} \hat{H}_i(\tau')$ satisfies:*

(1) $\bar{H}_i(\tau)$ is non-increasing in τ.

(2) $\bar{H}_i(\tau) \geq \hat{H}_i(\tau)$ for all τ.

(3) $\bar{H}_i(\tau) \leq 1$.

(4) $\bar{H}_i(\tau)$ is right-continuous in τ.

Proof. (1) If $\tau_1 < \tau_2$, then $\bar{H}_i(\tau_1) = \sup_{\tau' \geq \tau_1} \hat{H}_i(\tau') \geq \sup_{\tau' \geq \tau_2} \hat{H}_i(\tau') = \bar{H}_i(\tau_2)$. (2) Immediate. (3) The numerator of (9) is bounded above by $(1-\beta)\frac{1}{M}\sum_j \Phi^\tau(y_{ij} \mid x_i) \leq \hat{Z}^\tau(x_i, T_i)$, so $\hat{H}_i(\tau) \leq 1$, and a supremum of values in $[0,1]$ is in $[0,1]$. (4) Fix $\tau \in [0,1)$. Since there are finitely many particles, there is a $\tau^+ \in (\tau, 1]$ such that no score $S(y_{ij} \mid x_i)$ lies in (τ, τ^+). Then $\Phi^{\tau'}(y_{ij} \mid x_i) = \Phi^\tau(y_{ij} \mid x_i)$ for all j and all $\tau' \in [\tau, \tau^+)$, so $\hat{H}_i$ is constant on $[\tau, \tau^+)$. Hence, $\bar{H}_i = \sup_{\tau' \geq \cdot} \hat{H}_i(\tau')$ is right-continuous. □

Lemma B.5. *Let us refer to the test prompt x_{test} and tuple T_{test} as x_{n+1} and T_{n+1} respectively. Define the oracle*

$$\tau^\star = \inf\left\{\tau \mid \frac{1}{n+1}\sum_{i=1}^{n+1} \bar{H}_i(\tau) \leq \alpha\right\}. \tag{19}$$

(Note once again that this is well-posed: the feasible set at least contains $\tau = 1$, and by Lemma B.4 (1) and (4) the function is non-increasing and right-continuous in τ, so the infimum both exists and is attained.)

Then $\hat{\tau} \geq \tau^\star$ almost surely.

Proof. For any τ in the feasible set of $\hat{\tau}$, Algorithm 1 guarantees $\frac{1}{n+1}\sum_{i=1}^n \bar{H}_i(\tau) + \frac{1}{n+1} \leq \alpha$. Since by Lemma B.4, $\bar{H}_{n+1}(\tau) \leq 1$,

$$\frac{1}{n+1}\sum_{i=1}^{n+1} \bar{H}_i(\tau) \leq \frac{1}{n+1}\sum_{i=1}^{n} \bar{H}_i(\tau) + \frac{1}{n+1} \leq \alpha,$$

so τ is in the feasible set of $\tau^\star$, giving $\tau^\star \leq \hat{\tau}$ a.s. □

Main Proof of Theorem 3.2

The proof follows the general approach for conformal risk control proofs Angelopoulos et al. [2024a]

Proof. The test response is $y_{\text{test}} \sim \hat{p}^{\hat{\tau}}(\cdot \mid x_{\text{test}}, T_{\text{test}})$. Write $(x_{n+1}, T_{n+1}) := (x_{\text{test}}, T_{\text{test}})$ for the test tuple. Since the prompts $(x_i)_{i=1}^{n+1}$ are exchangeable and, conditional on these prompts, the particles in each T_i are drawn i.i.d. from $p_{\text{LLM}}(\cdot \mid x_i)$, the tuples $(x_i, T_i)_{i=1}^{n+1}$ are exchangeable as well.

Step 1: Empirical Posterior Risk at $\hat{\tau}$. When $y_{\text{test}} \sim \hat{p}^{\hat{\tau}}(\cdot \mid x_{\text{test}}, T_{\text{test}})$, Lemma B.3 applied with $\gamma = \hat{\tau}$ gives

$$\mathbb{E}[\ell(x_{\text{test}}, y_{\text{test}}) \mid x_{\text{test}}, T_{\text{test}}, \hat{\tau}] = \hat{H}_{n+1}(\hat{\tau}).$$

Combining with Lemma B.4 (2) and taking the outer expectation,

$$\mathbb{E}[\ell(x_{\text{test}}, y_{\text{test}})] = \mathbb{E}[\hat{H}_{n+1}(\hat{\tau})] \leq \mathbb{E}[\bar{H}_{n+1}(\hat{\tau})]. \tag{20}$$

Step 2: Transfer from $\hat{\tau}$ to $\tau^\star$. By Lemma B.5, $\hat{\tau} \geq \tau^\star$ a.s. Since $\bar{H}_{n+1}(\tau)$ is non-increasing in τ, by Lemma B.4,

$$\bar{H}_{n+1}(\hat{\tau}) \leq \bar{H}_{n+1}(\tau^\star), \quad \text{a.s.} \tag{21}$$

Step 3: Exchangeability. The tuples $(x_i, T_i)_{i=1}^{n+1}$ are exchangeable, and $\tau^\star = \tau^\star((x_i, T_i)_{i=1}^{n+1})$ is a symmetric function of these tuples. Conditioning on the multiset $E = \{(x_i, T_i)\}_{i \in [n+1]}$, exchangeability gives

$$\mathbb{E}[\bar{H}_{n+1}(\tau^\star) \mid E] = \frac{1}{n+1} \sum_{i=1}^{n+1} \bar{H}_i(\tau^\star) \leq \alpha, \quad \text{a.s.},$$

by the definition of $\tau^\star$ in (19). Integrating over E,

$$\mathbb{E}[\bar{H}_{n+1}(\tau^\star)] \leq \alpha. \tag{22}$$

Combining (20), (21), and (22) yields $\mathbb{E}[\ell(x_{\text{test}}, y_{\text{test}})] \leq \alpha$. □

B.3 Proof of Lemma 3.3

Before proving the lemma, we will need a technical lemma bounding the bracketing number van der Vaart and Wellner [1996] of some function classes. Let us denote briefly by $P_{Y,S|x}$ the conditional joint distribution of responses and scores given a prompt x.[14]

Lemma B.6. *Assume the distribution of the scores is continuous*[15]*. Let $h : \mathcal{V}^* \times \mathcal{S} \to \{0, 1\}$ be a fixed measurable map. Define the function class*

$$\mathcal{G} = \{(y, s) \mapsto h(y, s)\mathbf{1}\{s > \tau\} \mid \tau \in [0, 1]\},$$

and identify its elements as g_τ. Then, for every $\varepsilon > 0$,

$$N_{[]}\Big(\varepsilon, \mathcal{G}, \|\cdot\|_{L_1(P_{Y,S|x})}\Big) < \infty.$$

Proof. Let $P = P_{Y,S|x}$. For $\tau \in [0, 1]$, write

$$g_\tau(y, s) = h(y, s)\mathbf{1}\{s > \tau\}.$$

Fix $\tau_1 \leq \tau_2$. Then for every $\tau \in [\tau_1, \tau_2]$,

$$g_{\tau_2}(y, s) = h(y, s)\mathbf{1}\{s > \tau_2\} \leq h(y, s)\mathbf{1}\{s > \tau\} \leq h(y, s)\mathbf{1}\{s > \tau_1\} = g_{\tau_1}(y, s),$$

where we used that $h \geq 0$. Hence the bracket

$$[g_{\tau_2}, g_{\tau_1}]$$

contains all functions g_τ with $\tau \in [\tau_1, \tau_2]$.

The $L_1(P)$-width of this bracket is

$$\begin{aligned} \|g_{\tau_1} - g_{\tau_2}\|_{L_1(P)} &= \mathbb{E}_P\left[h(Y, S)\left(\mathbf{1}\{S > \tau_1\} - \mathbf{1}\{S > \tau_2\}\right)\right] \\ &= \mathbb{E}_P\left[h(Y, S)\mathbf{1}\{\tau_1 < S \leq \tau_2\}\right]. \end{aligned}$$

[14] We only model the joint distribution to handle randomized scorers; for deterministic scorers this is not necessary and a measure over $\mathcal{V}^*$ suffices.

[15] We enforce this by adding noise to the elicited scores.

Define the finite measure μ_h on $[0,1]$ by

$$\mu_h(A) = \mathbb{E}_P\left[h(Y,S)\mathbf{1}\left\{S \in A\right\}\right].$$

Since $h \in \{0,1\}$, we have $\mu_h([0,1]) \leq 1$. With this notation,

$$\|g_{\tau_1} - g_{\tau_2}\|_{L_1(P)} = \mu_h((\tau_1, \tau_2]).$$

Now fix $\varepsilon > 0$. Since μ_h is a finite measure on $[0,1]$, we can partition $(0,1]$ into finitely many intervals $I_k = (a_k, a_{k+1}]$ such that each interval has μ_h-mass at most ε. Since the scores are continuous, this measure has no atoms. For each such interval $I_k = (a_k, a_{k+1}]$, the bracket

$$[g_{a_{k+1}}, g_{a_k}]$$

has $L_1(P)$-width

$$\left\|g_{a_k} - g_{a_{k+1}}\right\|_{L_1(P)} = \mu_h((a_k, a_{k+1}]) \leq \varepsilon.$$

Thus finitely many $L_1(P)$-brackets of width at most ε cover $\mathcal{G}$. Therefore

$$N_{[]}\Big(\varepsilon, \mathcal{G}, \|\cdot\|_{L_1(P_{Y,S|x})}\Big) < \infty.$$

□

We are ready to prove Lemma 3.3.

Fix a prompt x and abstention mass $\beta > 0$. Write $T_M = (y_1, \ldots, y_M)$ with $y_j \overset{\text{i.i.d.}}{\sim} p_{\text{LLM}}(\cdot \,|\, x)$. Recall $\Phi^\tau(y \,|\, x) = \mathbf{1}\left\{S(y \,|\, x) > \tau\right\}$ where $S(y \,|\, x) = \prod_i s_i$ is the completion-level score. Define the quantities

$$\hat{A}_M(\tau) := \frac{1}{M}\sum_{j=1}^{M} \Phi^\tau(y_j \,|\, x)\, \ell(x, y_j), \qquad A(\tau) := \mathbb{E}_{y \sim p_{\text{LLM}}}[\Phi^\tau(y \,|\, x)\, \ell(x,y)],$$

$$\hat{B}_M(\tau) := \frac{1}{M}\sum_{j=1}^{M} \Phi^\tau(y_j \,|\, x), \qquad B(\tau) := \mathbb{E}_{y \sim p_{\text{LLM}}}[\Phi^\tau(y \,|\, x)].$$

The empirical and population mixture normalizers are

$$\hat{Z}_M^\tau = \beta + (1-\beta)\hat{B}_M(\tau), \qquad Z^\tau = \beta + (1-\beta)B(\tau),$$

both in $[\beta, 1]$. We dropped the dependence on x, as we fixed this variable already.

Step 1: Uniform convergence via Glivenko–Cantelli. Instantiating $h(y,s) = \ell(x,y)$ and $h(y,s) = 1$ shows that $\{y \mapsto 1\{S(y \mid x) > \tau\} : \tau \in [0,1]\}$ and $\{y \mapsto \ell(x,y)1\{S(y \mid x) > \tau\} : \tau \in [0,1]\}$ have finite Bracketing number, by Lemma B.6. By Theorem 2.4.1. in van der Vaart and Wellner [1996], both of these classes are therefore Glivenko-Cantelli. Therefore

$$\sup_{\tau \in [0,1]} \left|\hat{A}_M(\tau) - A(\tau)\right| \xrightarrow[M\to\infty]{\text{a.s.}} 0, \qquad \sup_{\tau \in [0,1]} \left|\hat{B}_M(\tau) - B(\tau)\right| \xrightarrow[M\to\infty]{\text{a.s.}} 0. \tag{23}$$

Step 2: Uniform convergence of the empirical-posterior risk. Recall that the empirical and population conditional risks are given by

$$\hat{R}_M = \mathbb{E}_{y \sim \hat{p}^\tau(\cdot \,|\, x, T_M)}[\ell(x,y)] \qquad \text{and} \qquad R = \mathbb{E}_{y \sim p^\tau(\cdot \,|\, x)}[\ell(x,y)].$$

This can be rewritten as

$$\hat{R}_M(x,\tau) = \frac{(1-\beta)\hat{A}_M(\tau)}{\hat{Z}_M^\tau}, \qquad R(x,\tau) = \frac{(1-\beta)A(\tau)}{Z^\tau}.$$

Adding and subtracting $(1-\beta)\hat{A}_M(\tau)/Z^\tau$, we get

$$\begin{aligned}
\left|\hat{R}_M(x,\tau) - R(x,\tau)\right| &= \left|\frac{(1-\beta)\hat{A}_M(\tau)}{\hat{Z}_M^\tau} - \frac{(1-\beta)A(\tau)}{Z^\tau}\right| \\
&= \left|\frac{(1-\beta)\hat{A}_M(\tau)}{\hat{Z}_M^\tau} - \frac{(1-\beta)\hat{A}_M(\tau)}{Z^\tau} + \frac{(1-\beta)\hat{A}_M(\tau)}{Z^\tau} - \frac{(1-\beta)A(\tau)}{Z^\tau}\right| \\
&\le \left|\frac{(1-\beta)\hat{A}_M(\tau)}{Z^\tau} - \frac{(1-\beta)A(\tau)}{Z^\tau}\right| + \left|\frac{(1-\beta)\hat{A}_M(\tau)}{\hat{Z}_M^\tau} - \frac{(1-\beta)\hat{A}_M(\tau)}{Z^\tau}\right| \\
&= \frac{1-\beta}{Z^\tau}\left|\hat{A}_M(\tau) - A(\tau)\right| + (1-\beta)\hat{A}_M(\tau)\left|\frac{1}{\hat{Z}_M^\tau} - \frac{1}{Z^\tau}\right| \\
&= \frac{1-\beta}{Z^\tau}\left|\hat{A}_M(\tau) - A(\tau)\right| + (1-\beta)\hat{A}_M(\tau)\frac{\left|\hat{Z}_M^\tau - Z^\tau\right|}{\hat{Z}_M^\tau Z^\tau}.
\end{aligned}$$

Since

$$\hat{Z}_M^\tau - Z^\tau = \left[\beta + (1-\beta)\hat{B}_M(\tau)\right] - \left[\beta + (1-\beta)B(\tau)\right] = (1-\beta)(\hat{B}_M(\tau) - B(\tau)),$$

and since $\hat{A}_M(\tau) \le 1$, $Z^\tau \ge \beta$, and $\hat{Z}_M^\tau Z^\tau \ge \beta^2$, it follows that

$$\left|\hat{R}_M(x,\tau) - R(x,\tau)\right| \le \frac{1-\beta}{\beta}\left|\hat{A}_M(\tau) - A(\tau)\right| + \frac{(1-\beta)^2}{\beta^2}\left|\hat{B}_M(\tau) - B(\tau)\right|.$$

Taking the supremum over $\tau \in [0,1]$ on both sides and applying (23),

$$\sup_{\tau\in[0,1]}\left|\hat{R}_M(x,\tau) - R(x,\tau)\right| \xrightarrow[M\to\infty]{\text{a.s.}} 0. \tag{24}$$

The same argument applies for P_X-a.e. x, completing the proof of Lemma 3.3.

B.4 Proof of Corollary 3.4

Since the guarantee above holds uniformly over τ, it in particular holds for the selected threshold $\hat{\tau}_M \in [0,1]$ from Algorithm 1 when run with M samples from p_{LLM}:

$$\left|\hat{R}_M(x_{\text{test}}, \hat{\tau}_M) - R(x_{\text{test}}, \hat{\tau}_M)\right| \le \sup_{\tau\in[0,1]}\left|\hat{R}_M(x_{\text{test}}, \tau) - R(x_{\text{test}}, \tau)\right|.$$

So by (24) and by the bounded convergence theorem,

$$\mathbb{E}\left[\sup_{\tau\in[0,1]}\left|\hat{R}_M(x_{\text{test}}, \tau) - R(x_{\text{test}}, \tau)\right|\right] \xrightarrow[M\to\infty]{} 0.$$

By Theorem 3.2,

$$\mathbb{E}[\hat{R}_M(x_{\text{test}}, \hat{\tau}_M)] \le \alpha.$$

Therefore, for $y_{\text{test}}^{\text{pop}} \sim p^{\hat{\tau}_M}(\cdot \mid x_{\text{test}})$,

$$\mathbb{E}[\ell(x_{\text{test}}, y_{\text{test}}^{\text{pop}})] = \mathbb{E}[R(x_{\text{test}}, \hat{\tau}_M)] \le \alpha + o(1),$$

and hence

$$\limsup_{M\to\infty} \mathbb{E}[\ell(x_{\text{test}}, y_{\text{test}}^{\text{pop}})] \le \alpha.$$

C Design Choices for the Conformal Routine

This appendix provides motivation and design rationale for the calibration rule in Section 3. The arguments here are not needed for the validity proofs of Proposition 3.1 and Theorem 3.2. Instead,

they explain how the rule arises from an idealized known-normalizer setting and why we use a fixed-denominator, monotonized objective. We highlight the connections to traditional approaches to conformal prediction with (fixed) covariate and label shift.

The calibration procedures developed in Section 3 are intuitively derived from the oracles τ^*_{id} and $\tau^\star$ in (16) and (19) respectively, both of which use a *fixed* denominator of $n+1$ and importance-reweight prior samples to evaluate/estimate the posterior risk. A natural alternative is to use an *adaptive* (self-normalized) denominator, as is common in the distribution-shift conformal prediction literature [Tibshirani et al., 2019, Angelopoulos et al., 2024b, Fannjiang et al., 2022, Prinster et al., 2024]. Fundamentally, this oracle arises as follows. Suppose that $U_i \sim P$ for $i \in [n]$ and $U_{n+1} \sim Q$. Then conditional on the multiset of observed values $E = \{U_1, \ldots, U_n, U_{n+1}\}$

$$U_{n+1} \,|\, E \sim \sum_{i=1}^{n+1} v_i \delta_{U_i} \quad \text{where} \quad v_i = \frac{\frac{dQ(U_i)}{dP(U_i)}}{\sum_{j=1}^{n+1} \frac{dQ(U_j)}{dP(U_j)}}.$$

This can be found in Proposition 7.6. of Angelopoulos et al. [2024b]. Extensions exist under Feedback covariate shift Fannjiang et al. [2022] and more arbitrary dependencies Prinster et al. [2024]. As a consequence, if we had $x_i \sim P_X$ for $i \in [n+1]$ and $y_{n+1} \sim p^\tau(\cdot \,|\, x_{n+1})$ for *fixed* τ, we'd have

$$(x_{n+1}, y_{n+1}) \,|\, E \sim \sum_{i=1}^{n+1} v_i \delta_{(x_i, y_i)} \quad \text{where } v_i = \frac{\frac{\Phi^\tau(y_i \,|\, x_i)}{Z^\tau(x_i)}}{\sum_{j=1}^{n+1} \frac{\Phi^\tau(y_j \,|\, x_j)}{Z^\tau(x_j)}}$$

In the known Z^τ case, this could suggest[16] an oracle selection rule

$$\tau^*_{\text{SN}} = \inf \left\{ \tau \,|\, \frac{\sum_{i=1}^{n+1} \frac{\Phi^\tau(y_i \,|\, x_i)}{Z^\tau(x_i)} \ell(x_i, y_i)}{\sum_{i=1}^{n+1} \frac{\Phi^\tau(y_i \,|\, x_i)}{Z^\tau(x_i)}} \leq \alpha \right\},$$

Intuitively, τ^*_{SN} computes a weighted average of the losses, with weights proportional to each point's likelihood under the posterior p^τ. One issue however, is that this is ill defined. For the weighted exchangeability to make sense, y_{n+1} now needs to be sampled according to p^τ. However, τ depends on y_{n+1}. This creates a circular dependency, that the choice of τ^*_{id} and $\tau^\star$ is able to sidestep, by depending only on a hypothetical calibration set of completions from the *prior*.

Population-Level Equivalence

We might still be curious about why τ^*_{id} and τ^*_{SN} are different, even though both seem principled. In fact, suppose all $n+1$ completions are drawn from the prior: $X_i \sim P_X$ and $Y_i \sim p(\cdot \mid X_i)$. Writing $w_i(\tau) = \Phi^\tau(Y_i \mid X_i)/Z^\tau(X_i)$, we have $\mathbb{E}[w_i(\tau) \mid X_i] = 1$ for any fixed τ, so the tower rule gives

$$\mathbb{E}\left[\sum_{i=1}^{n+1} w_i(\tau)\right] = \sum_{i=1}^{n+1} \mathbb{E}\left[\mathbb{E}\left[\frac{\Phi^\tau(Y_i \mid X_i)}{Z^\tau(X_i)} \,\middle|\, X_i\right]\right] = \sum_{i=1}^{n+1} \mathbb{E}\left[\frac{Z^\tau(X_i)}{Z^\tau(X_i)}\right] = n+1.$$

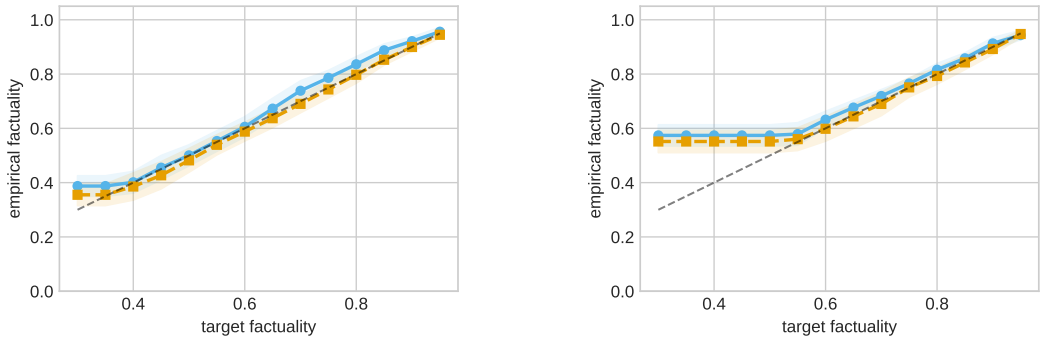
Thus, in expectation, the self-normalized denominator equals $n+1$. The two oracles can therefore be expected to behave similarly in large samples. Indeed, empirical experimentation has shown us that their resulting approximations behave very similarly in practice too. As our theoretical analysis employs τ^*_{id}, our Algorithm 1 also uses a fixed denominator.

D Additional Experimental Results

D.1 MATH Calibration Validity

Figure 4 shows the results of the empirical validity experiment (Figure 1) for both FActScore and MATH. MATH mirrors FActScore: empirically, valid coverage is achieved at all levels, with a flatlining region at targets low enough that both methods degrade to sampling from the prior.

[16] Ignoring for a second that monotonizing this objective is also trickier than if the denominator is fixed.

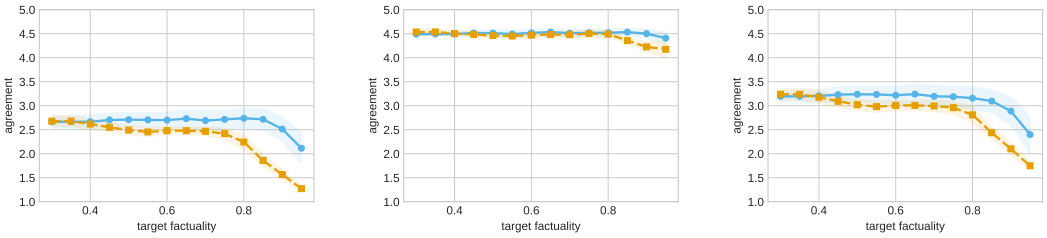


(a) FActScore. **(b)** MATH.

Figure 4. Target vs. observed factuality on FActScore (a) and MATH (b); ours (—●—) and MH (-■-), with $y = x$ (----). Both methods produce empirically valid coverage on both tasks. Averaged over 10 splits; shaded area shows 95% CIs.

D.2 Per-statement Likert Scores

Figure 5 breaks down the FActScore downstream utility results from Figure 2 across the three axes (see Appendix E.4). These results reveal that even at high targets, the LM judge considers non-abstaining outputs from both methods to be fluent and coherent. However, as the target factuality increases, outputs from MH quickly degrade in terms of their completeness and helpfulness. Note that Figure 2 is the arithmetic mean of these axes.



(a) Completeness | non-abstaining. **(b)** Fluency | non-abstaining. **(c)** Helpfulness | non-abstaining.

Figure 5. Breakdown of the FActScore Likert results from Figure 2 (panel b) per statement that the judge is asked to rate its agreement with, for ours (—●—) and MH (-■-): (a) whether the biography covers the most important facts a reader would expect to find in a short biography; (b) whether the biography reads as coherent prose with smooth transitions and no abrupt or disjointed jumps; and (c) whether this would be a satisfactory and helpful answer if a user asked for a short biography of the entity. The main-text Likert score is the per-output average over these three. Conditioned on non-abstention; averaged over 10 splits; shaded area shows 95% CI.

D.3 Human evaluation on FActScore

We additionally collected human utility ratings for generated biographies from one test split of 50 FActScore prompts at target factuality levels $0.6, 0.7, 0.8, 0.9$, using the same completeness, fluency, and helpfulness questions as the LM judges (App. E.4.1). Conditional on answering, our method receives higher mean ratings at every target; unconditionally, the two methods are close. Due to cost constraints, the evaluation covers a single split, whereas the LM-judge results throughout this paper average over ten splits. Participants were recruited through Prolific, provided informed consent, and were compensated at an average rate of $44/hour. The study received IRB approval[17]. In total, 200 annotators rated the 190 distinct responses that the two methods emit on this split (median five ratings per response); see Appendix F for details of the study.

To additionally validate that the LLM-as-a-judge evaluations reflect true differences in utility rather than biases of the particular judge, we also repeat the downstream evaluations with Claude Sonnet 5, with $\beta = 0.1$ as in the human study. Results are shown in Table 1: while raters differ in the

[17] In fact, the study qualified for an IRB exception.

Table 1. FActScore: mean Likert utility (1–5) by rater, unconditionally and conditioned on non-abstention, at $\beta = 0.1$, the setting of the human study. Human ratings are crowdsourced from 200 annotators (median 5 ratings per response) on a single test split of 50 prompts; automated judge results are averaged over 10 random splits. $\pm$ gives the half-width of a 95% bootstrap percentile interval, resampling the test prompts of each split (LLM judges) or the prompts and then each response's raters (humans).

		Target factuality			
Rater	Method	0.6	0.7	0.8	0.9
Unconditional					
GPT-5.4	Ours	3.05 $\pm$0.09	2.84 $\pm$0.10	2.62 $\pm$0.10	**2.26** $\pm$0.09
	MH	**3.20** $\pm$0.08	**3.10** $\pm$0.08	**2.91** $\pm$0.08	2.25 $\pm$0.06
Sonnet 5	Ours	2.09 $\pm$0.06	1.96 $\pm$0.06	1.81 $\pm$0.06	1.58 $\pm$0.06
	MH	**2.20** $\pm$0.05	**2.13** $\pm$0.06	**2.03** $\pm$0.05	**1.62** $\pm$0.04
Humans	Ours	**3.34** $\pm$0.30	3.08 $\pm$0.32	2.79 $\pm$0.34	2.47 $\pm$0.33
	MH	3.29 $\pm$0.29	**3.28** $\pm$0.31	**3.10** $\pm$0.30	**2.51** $\pm$0.28
Conditioned on non-abstention					
GPT-5.4	Ours	**3.48** $\pm$0.08	**3.50** $\pm$0.09	**3.48** $\pm$0.11	**3.34** $\pm$0.14
	MH	3.32 $\pm$0.08	3.31 $\pm$0.08	3.18 $\pm$0.08	2.63 $\pm$0.07
Sonnet 5	Ours	**2.37** $\pm$0.05	**2.38** $\pm$0.06	**2.36** $\pm$0.07	**2.27** $\pm$0.10
	MH	2.27 $\pm$0.05	2.27 $\pm$0.05	2.20 $\pm$0.05	1.87 $\pm$0.05
Humans	Ours	**3.69** $\pm$0.25	**3.72** $\pm$0.26	**3.79** $\pm$0.33	**3.76** $\pm$0.42
	MH	3.44 $\pm$0.28	3.46 $\pm$0.29	3.39 $\pm$0.28	3.02 $\pm$0.33

range of scores they assign (Sonnet 5 scoring lower than GPT-5.4 across the board, while the human annotators score slightly higher), all three raters show broadly consistent patterns in the utility differences between our method and MH. Conditional on answering, our method is preferred at every target by each judge, while unconditionally the two methods are closer and MH generally wins. Appendix D.4 repeats the LLM-as-judge comparison on MATH, where the conclusions are likewise unaffected by the choice of judge.

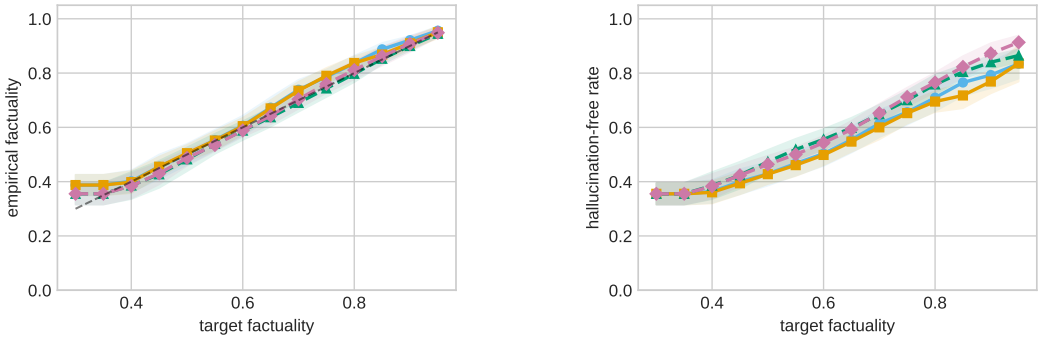
D.4 Judge Ablation on MATH

Complementing the FActScore rater comparison in Section D.3, we repeat the MATH completeness evaluation with Claude Sonnet 5 in place of GPT-5.4. Table 2 shows the results. The judges do sometimes disagree on individual responses, especially on heavily filtered MH outputs. For example, one MH response to the question "Solve for x: $\frac{x}{2} + \frac{x}{3} = 5$" reads, in full, "We verify: $(6/2) + (6/3) = 3 + 2 = 5$"; Sonnet 5 rates this complete, since the correct answer (6) is mentioned, whereas GPT-5.4 does not. However, in aggregate, the relative differences between our method and the baseline are similar under both judges.

Table 2. MATH: completeness rate (final answer correct) by judge model, unconditionally and conditioned on non-abstention. Mean over 10 seeds. ± gives the half-width of a 95% bootstrap percentile interval over the test prompts of each split.

		Target factuality			
Judge	Method	0.6	0.7	0.8	0.9
Unconditional					
GPT-5.4	Ours	**0.648** ±0.042	**0.602** ±0.042	**0.510** ±0.044	**0.354** ±0.042
	MH	0.540 ±0.044	0.240 ±0.037	0.128 ±0.030	0.046 ±0.018
Sonnet 5	Ours	**0.670** ±0.041	**0.614** ±0.042	**0.530** ±0.044	**0.360** ±0.041
	MH	0.604 ±0.042	0.318 ±0.041	0.186 ±0.034	0.080 ±0.024
Conditioned on non-abstention					
GPT-5.4	Ours	**0.719** ±0.042	**0.763** ±0.042	**0.808** ±0.044	**0.852** ±0.046
	MH	0.541 ±0.044	0.257 ±0.039	0.170 ±0.038	0.106 ±0.042
Sonnet 5	Ours	**0.743** ±0.040	**0.778** ±0.041	**0.843** ±0.040	**0.864** ±0.044
	MH	0.605 ±0.043	0.340 ±0.043	0.248 ±0.043	0.189 ±0.055

D.5 Alternative scoring functions and base models



(a) Target vs. observed factuality.

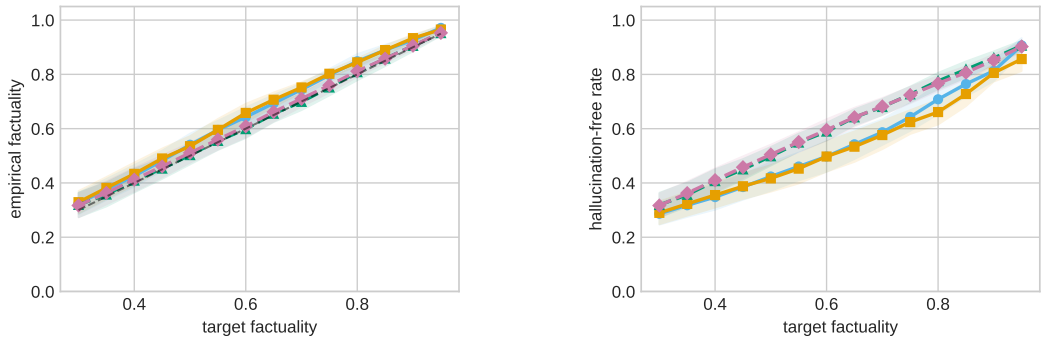
(b) Strict factuality of non-abstaining outputs.

Figure 6. Scorer-elicitation ablation on FActScore: confidence-rating elicitation (main text) vs. raw token log-probabilities. Ours with elicitation (—●—) and log-probability (—■—) scoring vs. MH with elicitation (- -▲- -) and log-probability (- -◆- -) scoring. Averaged over 10 splits; shaded area shows 95% CI.

The main text elicits per-claim scores for Llama-3.3-70B-Instruct generations by asking GPT-4o to rate its confidence on a 0-1 scale (Section 2.4). A natural, cheaper alternative is to read off the scorer's own token log-probabilities for a forced T/F/N/R judgment (true, false, neutral, refusal; prompt in Section E.3); we compute this scorer as $s = (p_{\mathtt{T}} + p_{\mathtt{N}} + p_{\mathtt{R}})/(p_{\mathtt{T}} + p_{\mathtt{N}} + p_{\mathtt{R}} + p_{\mathtt{F}})$ using GPT-4o. Figure 6 shows the impact that this alternative formulation has on calibration validity and sample utility. Both scorers produce empirically valid calibration curves, as predicted. Furthermore, for our method both scorers yield similar utilities; for MH, the logprob-based scorer yields slightly higher strict factuality among non-abstaining outputs at high targets in this experiment.

In Figure 7, we perform a related comparison: replacing both the base model and the scorer with smaller (and thus cheaper, but less capable) variants within the same family. Specifically, we gather generations from Llama-3.1-8B-Instruct [Grattafiori et al., 2024], which we then score using GPT-4o-mini instead of GPT-4o, using the (single-token and therefore cheaper) logprob-based MATH scorer of Section E.3 rather than the elicitation scorer of the main text. We fix the rest of the experimental setting as in Section 4.

As one can observe, validity is agnostic to the capacity of the base model and the scorer, while downstream utility can be affected somewhat, especially at higher target rates.



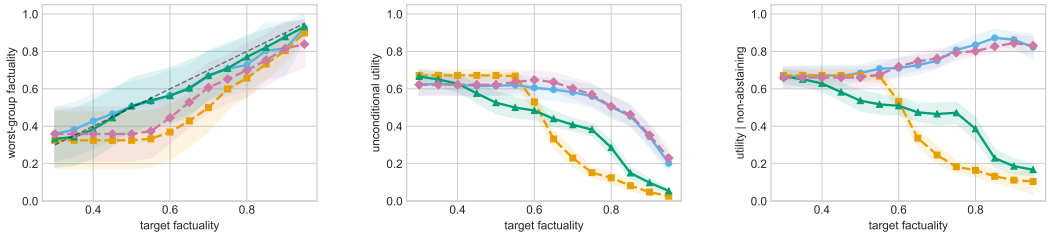
(a) Target vs. observed factuality.

(b) Strict factuality of non-abstaining outputs.

Figure 7. Model capacity ablation on MATH: GPT-4o vs. GPT-4o-mini logprob scoring, all on Llama-3.1-8B-Instruct generations (in contrast to the larger Llama-3.3-70B-Instruct used elsewhere). Ours with GPT-4o (—●—) and GPT-4o-mini (—■—) scoring vs. `MH` with GPT-4o (- -▲- -) and GPT-4o-mini (- -◆- -). Averaged over 10 splits; shaded area shows 95% CI.

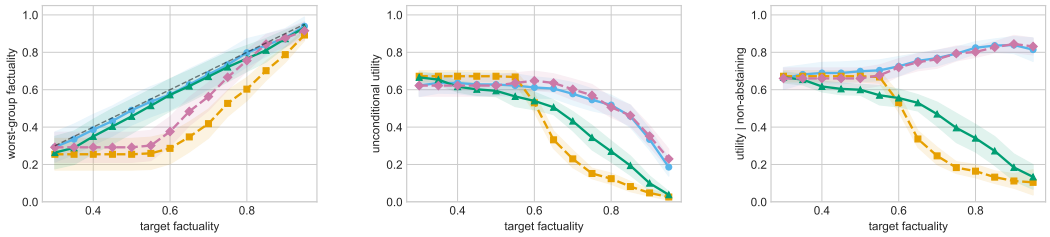
D.6 Group-conditional Factuality

Our posterior object can also be adapted to more sophisticated conformal guarantees, such as approximate conditional ones. We highlight this by running it separately per group, a method known (in the prediction set case) as Mondrian conformal prediction [Vovk et al., 2003]. In language modeling contexts, Cherian et al. [2024] have explored factuality under approximate conditional constraints. We refer to the method of Cherian et al. [2024] as `CGC`. We note that their method is more general and can be applied to various forms of conditional coverage, by viewing conditional coverage via a dual formulation over a restricted class of test functions. We instantiate it on the partition induced by the difficulties on the MATH dataset (of which there are 5), as well as the problem categories (of which there are 7). We use their randomized variant, designed to achieve the coverage budget exactly. Our Mondrian adaptation runs Algorithm 1 separately within each group G and samples from the empirical posterior using that group's threshold, guaranteeing that for any group $\mathbb{E}[\ell(x_{\text{test}}, y_{\text{test}}) \mid x_{\text{test}} \in G] \leq \alpha$. In Figures 8–11 we plot $\min_G \mathbb{E}[1 - \ell(x_{\text{test}}, y_{\text{test}}) \mid x_{\text{test}} \in G]$, as well as the resulting (marginal) unconditional utility and (marginal) conditional (on non-abstention) utility. Interestingly, we see that a group-conditional threshold has less of an effect on utility for our posterior calibration routine than for filtering approaches (where MH and the algorithm of Cherian et al. [2024]) differ more heavily for one of the scorer/base-model combinations, suggesting that our method already inherits a certain form of conditional adaptivity. Cherian et al. [2024]'s algorithm can improve utility over Mohri and Hashimoto [2024]'s algorithm because it can avoid too aggressive filtering for problems that are easier. Investigating this further is left for future work.



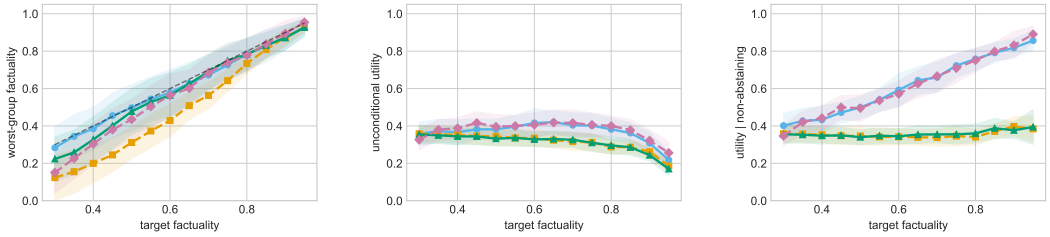
(a) Worst-group factuality. **(b)** Unconditional utility. **(c)** Utility | non-abstaining.

Figure 8. Llama-3.3-70B-Instruct with GPT-4o elicitation scoring, grouped by MATH's seven subjects. Ours (Mondrian; ——), ours (marginal; - - -), `CGC` (——), and `MH` (- - -), with $y = x$ (----) in (a). Averaged over 10 splits; shading shows approximate 95% intervals. For factuality (a), we bootstrap the splits and use the 95th percentile of the largest absolute change in any group mean as the interval half-width. Utility intervals (b, c) use t-intervals across splits, consistent with the rest of the paper.



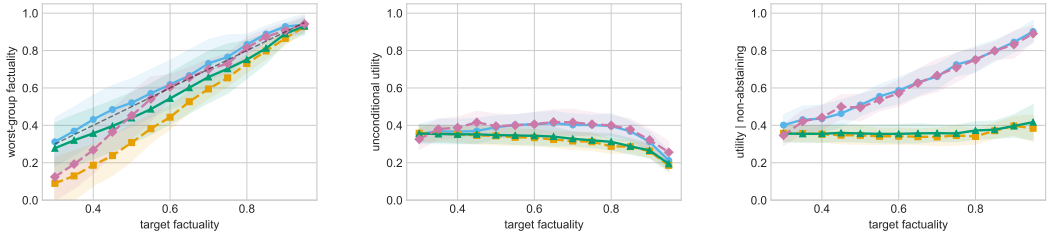
(a) Worst-group factuality. **(b)** Unconditional utility. **(c)** Utility | non-abstaining.

Figure 9. Llama-3.3-70B-Instruct with GPT-4o elicitation scoring, grouped by MATH's five difficulty levels. Ours (Mondrian; ——), ours (marginal; - - -), `CGC` (——), and `MH` (- - -), with $y = x$ (----) in (a). Averaged over 10 splits; shading shows approximate 95% intervals. For factuality (a), we bootstrap the splits and use the 95th percentile of the largest absolute change in any group mean as the interval half-width. Utility intervals (b, c) use t-intervals across splits, consistent with the rest of the paper.



(a) Worst-group factuality. **(b)** Unconditional utility. **(c)** Utility | non-abstaining.

Figure 10. Llama-3.1-8B-Instruct with GPT-4o logprob scoring, grouped by MATH's seven subjects. Ours (Mondrian; —●—), ours (marginal; - -◆- -), `CGC` (—▲—), and `MH` (- -■- -), with $y = x$ (----) in (a). Averaged over 10 splits; shading shows approximate 95% intervals. For factuality (a), we bootstrap the splits and use the 95th percentile of the largest absolute change in any group mean as the interval half-width. Utility intervals (b, c) use t-intervals across splits, consistent with the rest of the paper.



(a) Worst-group factuality. **(b)** Unconditional utility. **(c)** Utility | non-abstaining.

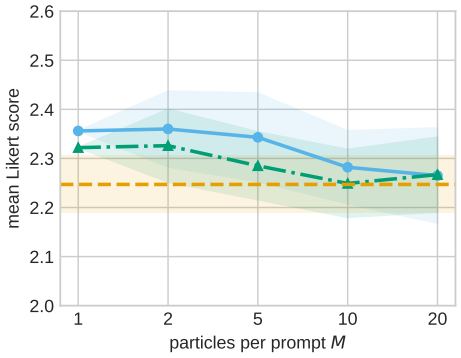
Figure 11. Llama-3.1-8B-Instruct with GPT-4o logprob scoring, grouped by MATH's five difficulty levels. Ours (Mondrian; —●—), ours (marginal; - -◆- -), `CGC` (—▲—), and `MH` (- -■- -), with $y = x$ (----) in (a). Averaged over 10 splits; shading shows approximate 95% intervals. For factuality (a), we bootstrap the splits and use the 95th percentile of the largest absolute change in any group mean as the interval half-width. Utility intervals (b, c) use t-intervals across splits, consistent with the rest of the paper.

D.7 The Impact of the Particle Budget M on Downstream Utility

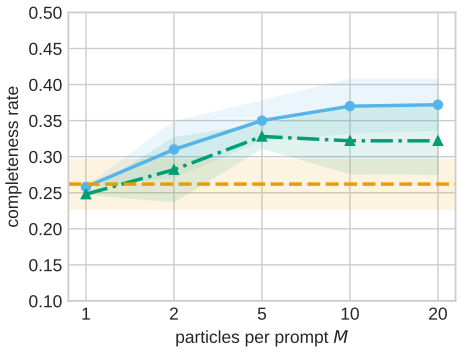
In our main experiments, we form the empirical posterior using $M = 20$ draws per prompt. However, gains from our method in this case confound two changes compared to claim-level `MH`: filtering on the output-level instead of on the claim-level and benefiting from increased test-time compute (that is, drawing more particles per prompt). One particularly illustrative comparison is to look at $M = 1$ with $\beta \to 0$ (which we implement as $\beta = 10^{-5}$ in practice). In this case the empirical posterior consists of a single particle and almost surely abstains if and only if that particle is rejected, essentially making our method an *output-level* analogue of the claim-level `MH`. To investigate when gains arise from output-level abstentions versus from increasing test-time compute, we therefore repeat our experiments with varying M, keeping the rest of the setup as in Section 4. As we vary M, we rerun calibration for each particle budget, using M particles per prompt for both calibration and evaluation.

Figure 12a shows the results on FActScore, using the elicitation scorer of the main experiments. Counter-intuitively, in this setting the mean Likert score decreases with M (2.36 at $M = 1$ against 2.27 at $M = 20$ for $\beta \to 0$). This suggests that the differences to `MH` reported in Section 4.2 arise primarily from filtering at the level of whole outputs, which concentrates abstentions on the prompts that the model struggles most often with, rather than from increased test-time compute.

There are a few possible reasons why test-time scaling may not lead to the expected utility gains in the FActScore domain. One is that the judge does not necessarily reward truthfulness, the loss we control. Maintaining risk control as M grows requires a stricter threshold, which biases towards short, conservative biographies; the judge likely rates these as less complete and helpful even if they are in fact more factual. Another potential reason why test-time scaling is not effective in this domain is that our rather cautious generation prompt (Section E.1) explicitly instructs the model not to speculate; for biographical generation, this perhaps leaves too little variation in the nature of the responses for test-time scaling to be effective.



(a) FActScore, Llama-3.3-70B-Instruct, GPT-4o elicitation scorer.



(b) MATH, Llama-3.1-8B-Instruct, GPT-4o logprob scorer.

Figure 12. Judged utility as a function of the particle budget M at target factuality 0.9, with abstentions scored as in Section 4.2; ours at $\beta \to 0$ (—●—) and $\beta = 0.05$ (-·▲·-), claim-level `MH` (----, independent of M). Shaded bands are paired 95% intervals for the difference to $M = 1$ over 10 splits. Validity holds at every point. Note the reduced range of the y-axis.

However, these results do not falsify the intuition that our procedure can enable provably safe test-time scaling when the generating model, scorer and downstream measure of utility are adequately matched. Figure 12b repeats the sweep on MATH with Llama-3.1-8B-Instruct generations scored by the GPT-4o logprob scorer of Section D.5. In this case, we observe that completeness at target 0.9 increases monotonically from 0.26 at $M = 1$ to 0.37 at $M = 20$ for $\beta \to 0$, significantly so at every $M > 1$, while remaining empirically valid throughout. At $\beta = 0.05$ the gain shrinks to about two thirds (0.25 to 0.32), but nonetheless ends up beating out `MH`. This shows that in a setting where the truthfulness of individual claims aligns with downstream utility and the scorer is able to separate the signal from the noise across the generated samples, calibrating for posterior sampling is an effective way of obtaining both risk control and test-time scaling.

Finally, to further isolate the effect that the scorer has on test-time scaling, we interpolate between this scorer and a perfect one by synthetically sharpening the scorer's predictions in the direction of each claim's true label. Specifically, we replace each claim's score s by $\sigma(\mathrm{logit}(s) + \delta)$ if the claim is true and $\sigma(\mathrm{logit}(s) - \delta)$ if it is false, where σ is the logistic function and $\delta > 0$ is the amount to shift by. The results are shown in Figure 13; notably, as the scorer is sharpened, `MH` (which cannot benefit from additional draws due to its single-draw filtering objective) only improves from 0.25 to 0.37 ($+0.12$), whereas our method at $M = 20$ improves from 0.37 to 0.59 ($+0.22$). Meanwhile, at $M = 1$, our method at best tracks per-claim `MH`, and even slightly underperforms it as the scorer gets better. These findings thus validate the intuition that as the scorer gets increasingly proficient at distinguishing true and false claims (rather than acting as an increasingly capable prompt-level difficulty indicator), our method enables larger and larger utility gains through increased test-time compute (while guaranteeing risk control, unlike naïve test-time methods) rather than through output-level filtering alone.

D.8 Ablating β at $M = 20$ with fixed-budget rejection sampling

This appendix probes how our asymptotic guarantee for rejection sampling (Lemma 3.3 and Corollary 3.4) interacts with β and a fixed-budget rejection sampling deployment procedure. We fix $M = 20$ as in the main experiments, but at deployment we run rejection sampling up to a hard cap of 50 attempts; if no draw (a generation or abstention) is accepted within this budget, we return a fallback sample. We consider two fallback choices that intuitively give upper and lower bounds on the unknown behavior of the true mixture posterior:

- **Worst-case fallback**: a synthetic sample consisting of a single claim which is incorrect but has a score 1.0, contributing maximally to the risk estimate. This deliberately pessimistic view assumes that every budget-exhausted prompt would produce a fully false, accepted output.
- **Abstention fallback**: when the budget is exhausted, the `abstain` response is deterministically sampled. Since this contributes zero to the loss/risk estimate, this is an optimistic

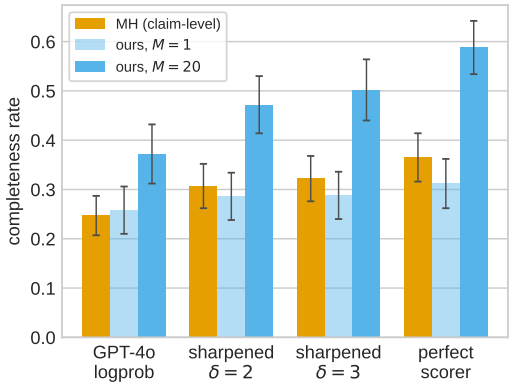
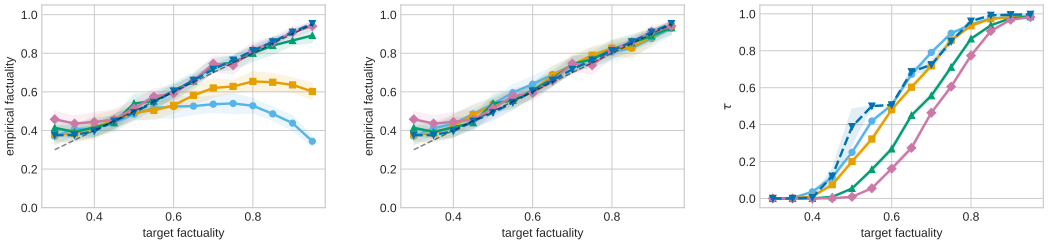


Figure 13. Judged completeness on MATH (Llama-3.1-8B-Instruct, target 0.9, $\beta \to 0$) as the scorer is sharpened toward the claim labels. The two synthetic scorers are obtained by shifting the GPT-4o logprob scorer's logits by $\delta \in \{2, 3\}$ toward the label, while the perfect scorer is simulated by setting the score to the true label plus/minus $|\mathcal{N}(0, 0.01^2)|$ noise. The error bars show 95% intervals over 10 splits. Note the reduced range of the y-axis.

view motivated by the observation that the mixture prior would have eventually emitted an abstention had we kept drawing. This fallback would also likely be the choice of practitioners using budget-constrained sampling techniques, especially in deployment settings where abstention is favored over returning a possibly incorrect generation.

We contrast these two fallbacks as the true behavior under indefinite rejection sampling lies somewhere between the two. By comparing the two, we can thus get a sense of how far the asymptotic guarantee of Corollary 3.4 can be from being empirically valid at $M = 20$ under different values of β. However, note that fully characterizing the precise support of the posterior is not possible with finite sample sizes; these fixed-budget fallbacks only give a coarse and loose approximation.



(a) Target vs. observed factuality (worst-case fallback).

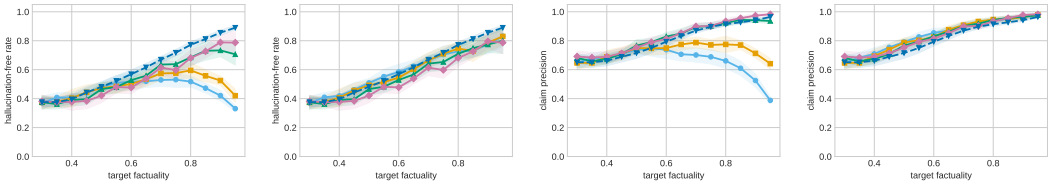
(b) Target vs. observed factuality (abstention fallback).

(c) Calibrated threshold $\hat{\tau}$ (does not depend on the fallback).

Figure 14. Fixed-budget rejection-sampling β-ablation on FActScore. (a, b) Target vs. empirical factuality under the worst-case (fully-false) and abstention fallbacks; (c) Calibrated $\hat{\tau}$ vs. target (identical for both fallbacks, since calibration is unaffected). Ours at $\beta = 0.001$ (—●—), $\beta = 0.01$ (—■—), $\beta = 0.05$ (—▲—) and $\beta = 0.1$ (—◆—), vs. `MH` (- -▼- -). Averaged over 10 splits; shaded area shows 95% CI.

Calibration validity. Figures 14a and 14b show target factuality $1 - \alpha$ against observed factuality on the held-out test split for different values of β, under the worst-case and abstention fallbacks respectively. At a relatively low calibration budget of $M = 20$, as $\beta \downarrow 0$ the calibration guarantee begins to fall apart under the worst-case fallback, yielding undercalibrated responses at high targets due to an increasing fraction of prompts hitting the budget cap; at $\beta = 0.1$, coverage is empirically valid and tight even under this most pessimistic scenario, while at our default $\beta = 0.05$ it falls slightly short only at the two highest targets (0.87 at 0.9 and 0.89 at 0.95). These correspond to prompts for which the language model has a hard time generating responses that pass the scorer threshold. We conjecture that with a reliable scorer, we would eventually return an abstention particle, but this is impossible to verify under this sampling design. Under the abstention fallback, by contrast, coverage holds across the whole range of β.

Proxy metrics. Figure 15 reports two factuality proxy metrics from the main experiments, evaluated under the fixed-budget rejection-sampling procedure rather than the analytic plug-in, again under



(a) Strict fact. | non-abst. (worst-case). **(b)** Strict fact. | non-abst. (abst. fallback). **(c)** Claim precision (worst-case). **(d)** Claim precision (abst. fallback).

Figure 15. Proxy metrics under the fixed-budget rejection sampling procedure on FActScore, comparing the worst-case (fully-false) and `abstention` budget-exhausted fallbacks. Ours at $\beta = 0.001$ (—●—), $\beta = 0.01$ (—■—), $\beta = 0.05$ (—▲—) and $\beta = 0.1$ (—◆—), vs. `MH` (--▼--).

both fallbacks. Under the worst-case fallback, the proxy metrics degrade as calibration fails at low β, mirroring the calibration plots. Under the abstention fallback, both strict factuality among non-abstaining samples and claim precision remain high across the whole range of β, since the budget-exhausted prompts do not affect the metrics. Generally speaking, except for in the instances of worst-case fallback-induced miscalibration, lower β leads to a closer gap in strict factuality among non-abstaining samples; this is because a lower forced probability of abstention in the mixture prior yields higher values of $\hat{\tau}$ (Figure 14c), thus applying stricter checks on the outputs.

D.9 `MH` with Reconstruction

As discussed in Section 4, our re-implementation of Mohri and Hashimoto [2024] differs from their original pipeline in two respects. First, we decompose responses into claims symbolically, by prompting the generator to write atomic sentences and splitting on sentence boundaries, rather than asking an external language model to extract claims. Second, we therefore omit their final *reconstruction* step, in which an external language model re-assembles the retained claims into a coherent generation. The scoring, calibration, and filtering mechanisms all match the original pipeline. Here we ablate the omission of reconstruction.

Reconstruction. We run `MH` with reconstruction, following Mohri and Hashimoto [2024] exactly: retained claims are merged by GPT-4o using the "merger" prompts from their Appendix A. Tables 3 and 4 show the results. On FActScore, merging does not noticeably improve utility. On MATH, it improves the completeness rate at higher targets, but the gap to our method remains significant across the whole range of targets. Reconstruction is not free, however. Inspecting the reconstruction step on 20 random responses filtered at a target of 0.9, we found that in 5 out of the 20 cases it added new reasoning steps. On inputs where only a claim or two survive the filter, the merger sometimes reconstructs the entire derivation on its own; the completeness gains on MATH thus partly reflect the merger model re-solving the problem, rather than the risk-controlled system itself. One such case is shown below: two elementary logarithm facts survive the filter, and the merger, despite its prompt instructing it to use only the steps provided, produces a complete change-of-base argument, derives the telescoping product, and arrives at the final answer on its own. This contradicts the intuitive meaning of Assumption 5.1 in Mohri and Hashimoto [2024] (namely that reconstruction neither removes nor introduces claims) and suggests that reconstruction improves completeness at the cost of provable risk control. Under our setup, where decomposition is symbolic and no reconstruction is used, the assumption holds trivially.

Problem:

> Find the value of $\log_2 3 \cdot \log_3 4 \cdot \log_4 5 \cdot \log_5 6 \cdot \log_6 7 \cdot \log_7 8$.

Surviving claims fed to the merger:

> We know that $8 = 2^3$, so $\log 8 = \log 2^3$.
> Using the property of logarithms, we have $\log 2^3 = 3 \log 2$.

Merged output:

> To solve the problem, we can use the given steps and properties of logarithms:

1. We know that $8 = 2^3$, so $\log 8 = \log 2^3$.
2. Using the property of logarithms, we have $\log 2^3 = 3 \log 2$.
Now, let's apply these steps to the problem:
The expression we need to evaluate is:

$$\log_2 3 \cdot \log_3 4 \cdot \log_4 5 \cdot \log_5 6 \cdot \log_6 7 \cdot \log_7 8$$

Using the change of base formula, each term can be rewritten as:

$$\log_2 3 = \frac{\log 3}{\log 2}, \quad \log_3 4 = \frac{\log 4}{\log 3}, \quad \log_4 5 = \frac{\log 5}{\log 4},$$
$$\log_5 6 = \frac{\log 6}{\log 5}, \quad \log_6 7 = \frac{\log 7}{\log 6}, \quad \log_7 8 = \frac{\log 8}{\log 7}$$

Substituting these into the expression, we have:

$$\frac{\log 3}{\log 2} \cdot \frac{\log 4}{\log 3} \cdot \frac{\log 5}{\log 4} \cdot \frac{\log 6}{\log 5} \cdot \frac{\log 7}{\log 6} \cdot \frac{\log 8}{\log 7}$$

Notice that all intermediate terms cancel out, leaving:

$$\frac{\log 8}{\log 2}$$

From step 1 and step 2, we know:

$$\log 8 = 3 \log 2$$

Thus, the expression simplifies to:

$$\frac{3 \log 2}{\log 2} = 3$$

Therefore, the value of the expression is $\boxed{3}$.

Table 3. FActScore: mean Likert utility (1–5, GPT-5.4 judge) for the `MH` ablations. "`MH` + reconstruction" adds Mohri and Hashimoto [2024]'s language model merger step. As in the main experiments, we use `Llama-3.3-70B-Instruct` with GPT-4o elicitation scoring. Mean over 10 random splits of the data; ± shows the 95% confidence interval over the splits.

	Target factuality			
Method	0.6	0.7	0.8	0.9
Unconditional				
Ours	3.06 ±0.11	2.84 ±0.13	2.67 ±0.11	2.27 ±0.07
`MH`	3.20 ±0.08	3.10 ±0.09	2.91 ±0.08	2.25 ±0.06
`MH` + reconstruction	3.18 ±0.10	3.07 ±0.11	2.91 ±0.09	2.33 ±0.06
Conditioned on non-abstention				
Ours	3.48 ±0.14	3.47 ±0.17	3.47 ±0.17	3.30 ±0.21
`MH`	3.32 ±0.06	3.31 ±0.07	3.18 ±0.11	2.63 ±0.08
`MH` + reconstruction	3.30 ±0.09	3.28 ±0.09	3.18 ±0.12	2.76 ±0.06

Table 4. MATH: completeness rate (final answer correct, GPT-5.4 judge) for the `MH` ablations, columns as in Table 3. As in the main experiments, we use `Llama-3.3-70B-Instruct` with GPT-4o elicitation scoring. Mean over 10 random splits of the data; ± shows the 95% confidence interval over the splits.

	Target factuality			
Method	0.6	0.7	0.8	0.9
Unconditional				
Ours	0.648 ±0.045	0.602 ±0.055	0.510 ±0.055	0.354 ±0.054
`MH`	0.540 ±0.059	0.240 ±0.042	0.128 ±0.027	0.046 ±0.017
`MH` + reconstruction	0.600 ±0.042	0.346 ±0.049	0.218 ±0.048	0.088 ±0.024
Conditioned on non-abstention				
Ours	0.719 ±0.039	0.763 ±0.046	0.808 ±0.033	0.852 ±0.049
`MH`	0.541 ±0.058	0.257 ±0.042	0.170 ±0.031	0.106 ±0.035
`MH` + reconstruction	0.601 ±0.041	0.371 ±0.046	0.291 ±0.057	0.205 ±0.044

D.10 Abstention Rates

Table 5 reports how often each method abstains as the target level increases. For our method, we report the mean posterior abstention mass $\hat{\beta}$. For `MH`, we report the fraction of responses from which every claim was filtered, which we treat as implicit abstentions.

We find that our method abstains more often than `MH` on FActScore. On MATH, `MH`'s implicit abstention rate rises steeply with the target, exceeding our method's at the highest target level. As to the impact of the hyper-parameters, increasing the prior abstention mass to $\beta = 0.05$ translates to at most about 0.03 extra abstention mass in the posterior, at every target on both datasets. The particle budget M similarly has a small impact on the abstention rates: at $\beta \to 0$, drawing $M = 20$ particles instead of one raises the rate on MATH by between 0.04 and 0.08 depending on the target, while on FActScore the small decrease it produces is within the noise of our ten splits at every target. Intuitively, increasing M leads to two opposing effects on the abstention rate. On the one hand, drawing more particles gives more opportunities for one of them to be accepted. On the other hand, maintaining risk control with a larger particle budget requires a stricter calibrated threshold, increasing the rate at which individual samples are rejected. What these results thus suggest is that these two opposing factors mostly cancel out in practice, making M have little effect on the overall abstention rates; however, this naturally depends heavily on the exact generator and scorer used. Either way, regardless of the abstention rates, risk control is of course guaranteed by either procedure.

Table 5. Average abstention rates as the target level increases. For our method, we report the mean posterior abstention mass $\hat{\beta}$; for `MH`, we report the fraction of responses which are entirely empty. Mean over 10 random splits of the data; $\pm$ shows the 95% confidence interval over the splits.

		Target factuality			
Method	M	0.6	0.7	0.8	0.9
FActScore					
Ours, $\beta \to 0$	1	0.240 $\pm$0.038	0.334 $\pm$0.036	0.476 $\pm$0.046	0.612 $\pm$0.043
	20	0.204 $\pm$0.034	0.296 $\pm$0.028	0.428 $\pm$0.049	0.588 $\pm$0.041
Ours, $\beta = 0.05$	1	0.259 $\pm$0.038	0.352 $\pm$0.035	0.493 $\pm$0.043	0.630 $\pm$0.042
	20	0.205 $\pm$0.026	0.322 $\pm$0.033	0.433 $\pm$0.026	0.609 $\pm$0.034
`MH`	—	0.072 $\pm$0.021	0.119 $\pm$0.030	0.160 $\pm$0.034	0.373 $\pm$0.028
MATH					
Ours, $\beta \to 0$	1	0.048 $\pm$0.019	0.152 $\pm$0.022	0.294 $\pm$0.036	0.476 $\pm$0.055
	20	0.086 $\pm$0.033	0.206 $\pm$0.042	0.332 $\pm$0.062	0.552 $\pm$0.074
Ours, $\beta = 0.05$	1	0.075 $\pm$0.014	0.185 $\pm$0.027	0.310 $\pm$0.029	0.491 $\pm$0.050
	20	0.092 $\pm$0.011	0.193 $\pm$0.027	0.341 $\pm$0.045	0.562 $\pm$0.055
`MH`	—	0.002 $\pm$0.002	0.076 $\pm$0.019	0.269 $\pm$0.035	0.594 $\pm$0.046

D.11 Qualitative Examples

In this section we show three examples from FActScore, drawn at a target factuality of 0.8. For the first two, we show a non-abstaining response from our method next to `MH`'s filtered generation with the same seed. The final example shows how abstentions are formatted in the downstream evaluations.

Entity: Isla Fisher. *Ours:*

> Isla Fisher is an Australian actress. She was born on February 3, 1976. Isla Fisher is known for her roles in comedic films. She married comedian Sacha Baron Cohen in 2010.

`MH`:

> She was born on February 3, 1976. Isla Fisher is known for her roles in comedic films. She married comedian Sacha Baron Cohen in 2010.

Entity: Jagan. *Ours:*

Jagan is an Indian film actor. He primarily works in the Tamil film industry. Jagan has appeared in numerous Tamil films as a character actor. He is known for his supporting roles in various movies.

MH:

Jagan is an Indian film actor. Jagan has appeared in numerous Tamil films as a character actor. He is known for his supporting roles in various movies.

Entity: Andreea Marin. *Ours:*

I don't know who or what "Andreea Marin" is.

MH:

She was born on December 22, 1974.

E Prompts

E.1 Generation prompts

We do not set a special system prompt for generation; the dataset-specific prompts below are passed as the user's first turn.

FActScore.

You are a helpful assistant that writes biographies. You will be given the name of a well-known entity, such as a historical figure, scientist, artist, or music group. Write a concise biography of that entity. Include only facts you are confident are correct, and omit details you are uncertain about. Brief biographies are fine – do not pad with guesses. Keep each sentence focused on a single fact. Prefer short, atomic sentences over long sentences with multiple clauses.

Example:
Tell me a bio of Albert Einstein.
Albert Einstein was a famed physicist. He was born in Germany in 1879. Einstein is best known for developing the theory of relativity. He published the equation E=mc^2 in 1905. In 1921, he won the Nobel Prize in Physics. He died in 1955.

Now:
Tell me a bio of {entity}.

Where the *entity* is the entity's name, as given in the dataset.

MATH.

You are a mathematical reasoning assistant. Solve problems step by step in a concise manner. IMPORTANT: Use exactly one sentence per ATOMIC logical reasoning step, NEVER chain more than one step in a sentence. Only include reasoning steps which you are confident are correct.

Example:
Problem: Find the value of x such that $\sqrt{x+7} = 9$.
Solution:
We need to solve for x where the square root of (x + 7) equals 9.
Squaring both sides gives x + 7 = 81.
Subtracting 7 from both sides yields x = 74.
We verify: sqrt(74 + 7) = sqrt(81) = 9.
Therefore, x = 74.

Now solve:
Problem: `{raw_prompt}`
Solution:

Where the *raw_prompt* is the textual specification of the problem, as given in the dataset.

E.2 Data labeling prompts

FActScore. The biography is segmented into one claim per line and the entity's Wikipedia article is included in-context as the source of truth.

> [TASK DESCRIPTION]
> You are a fact-checking assistant, assessing a biography written about a notable person or entity. You will be supplied the name of the entity, a Wikipedia article about that entity, and a biography. The biography consists of a series of claims, conversational remarks or other statements, given line by line.
> Your task is to classify each line following these guidelines:
>
> - **true**: The line makes one or more factual claims about the entity, and EVERY claim is supported by the Wikipedia article.
> - **false**: The line makes one or more factual claim about the entity, but AT LEAST ONE claim is contradicted by the Wikipedia article.
> - **neutral**: The line is purely conversational or otherwise does not make any factual claims about the requested entity.
> - **refusal**: The line professes ignorance about the entity on the speaker's part, it expresses that the speaker is uncertain about who the name might be referring to, or the speaker is indicating that they do not have enough information to say which one out of multiple entities is being referred to.
>
> IMPORTANT: All lines containing factual claims about the entity MUST be evaluated as either 'true' or 'false'.
> [ENTITY] `{entity}`
> [WIKIPEDIA ARTICLE] `{wikipedia_content}`
> [BIOGRAPHY TO EVALUATE] `{claims_text}`
> [RESPONSE FORMAT] You must return an answer which follows the given JSON schema, with exactly `N` annotations, one for each line in order.

MATH.

> You are evaluating a mathematical proof/solution.
> Problem: `{problem}`
> Reference solution: `{reference_solution}`
> Generated solution steps to annotate (claims start with `[CLAIM #]`):
> `{claims_text}`
> For each step, classify as:
>
> - **true**: If the step is a mathematically correct derivation, *assuming* the previous steps are correct. Claims can be true without significantly advancing the proof.
> - **false**: If the step contains an error, or contradicts the previous steps.
> - **neutral**: use for neutral claims. For instance when all of the claims in the output together are expressing uncertainty or inability to solve the problem (e.g. "I am not able to solve…"), label each one of them as neutral. Use neutral too for steps that are exclusively filler text (e.g. "Let me answer this for you…").
>
> Also determine:
>
> - **complete**: Does the proof reach a final answer AND is the final answer correct (matches reference)?
>
> IMPORTANT: You MUST return exactly `N` annotations, one for each step in order.

E.3 Score function prompts

We here list the prompts corresponding to the *full-response elicitation* strategy, that is the strategy described in Section 2.4. We also give the prompts used for the alternative logprob-based score functions used in Appendix D.5 (FActScore) and Figure 7 (MATH).

FActScore (full-response elicitation).

> You will be given a context (e.g., a question) and a statement made in response to that question. Your job is to rate this statement on a scale from 0.0 to 1.0. The score should represent your confidence in the claims made by the statement, where a 1 is obvious facts

and results like ‘The earth is round’ and ‘1+1=2’. A 0 is for claims that are objectively false, like “the earth is flat”. On the other hand, if you’re unsure, you should assign probabilities that are strictly in between 0 and 1. If the statement is not a factual claim (e.g., “I don’t know”, “I cannot answer that”, “I’m sorry but…”, questions, or expressions of uncertainty), assign a score of 1.0 since these are not false claims. You may think out loud first, but you MUST end your response with “Score: X.XX” where X.XX is a number between 0.0 and 1.0.

The context:
`{prompt}`

Previously accepted facts (assume these are correct):
`{prior_claims}`

The statement:
`{new_claim}`

Your response:

MATH (full-response elicitation).

You are evaluating a step in a mathematical derivation.

You are given:

1. A math problem
2. Previous reasoning steps S
3. A new step C

Give your best estimate of the probability that the new step C is true, using S as context. In other words, you should estimate P(C correct | S). Rate your confidence on a scale from 0.00 to 1.00, and make sure your uncertainty estimates are calibrated, meaning, that the average rate of correctness should correspond to your prediction, conditional on your prediction. Reserve 1.00 for claims you are certain are correct, except for:

- If the step is not a mathematical claim (e.g., “Let me think…”, “I’ll solve this by…”), assign 1.00.
- Similarly, if the step is an abstention (e.g. “I do not know how to solve this”), assign 1.00.

Important: return confidence to two decimal digits!

Think briefly, then end with “Score: X.XX” where X.XX is between 0.00 and 1.00.

Math Problem:
`{prompt}`

Previous steps (assume correct):
`{prior_claims}`

New step to evaluate:
`{new_claim}`

Your evaluation:

FActScore (logprob scorer). The score is $(p_{\mathtt{T}} + p_{\mathtt{N}} + p_{\mathtt{R}})/(p_{\mathtt{T}} + p_{\mathtt{N}} + p_{\mathtt{R}} + p_{\mathtt{F}})$.

You are evaluating a snippet from a biography about a notable person or entity. The biography consists of a series of claims, conversational remarks or other statements, given snippet by snippet. You will be given the preceding snippet, which should be assumed to be correct, and a new snippet to evaluate.

Your task is to classify the new snippet following these guidelines:

- **T** = True: The snippet makes one or more factual claims about the requested entity, and you believe EVERY claim is likely to be correct.
- **F** = False: The snippet makes one or more factual claims about the requested entity, but you believe AT LEAST ONE claim is possibly incorrect.
- **N** = Neutral: The snippet is purely conversational or otherwise does not make any factual claims about the requested entity.
- **R** = Refusal: The snippet professes ignorance about the entity on the speaker’s part, it expresses that the speaker is uncertain about who the name might be referring to, or the speaker is indicating that they do not have enough information to say which one out of multiple entities is being referred to.

Claims which are obvious facts and results like "The earth is round", "1+1=2", and well-known facts about notable people should be classified as T. Claims about things that are very obscure or difficult for anyone to know, like the birthdays or other specific facts about non-notable people, should be classified as F.

IMPORTANT: Reply ONLY with exactly one letter: T, F, N, R. Do not reply with anything other than one of these letters.

Biography subject:
`{entity}`

Previous statements made by the speaker (assume these to be correct):
`{prior_claims}`

Snippet to evaluate:
`{new_claim}`

MATH (logprob scorer). The score is $(p_{\texttt{Y}} + p_{\texttt{C}})/(p_{\texttt{Y}} + p_{\texttt{N}} + p_{\texttt{C}})$.

You are evaluating a step in a mathematical derivation. Reply with only a single letter. You will get a list of prior claims, and a new step. Reply:

- **Y** = Yes: The step is mathematically valid given the prior steps.
- **N** = No: The step contains an error or contradicts the previous steps.
- **C** = Conversation: the step is not a mathematical claim (e.g., "Let me think about this carefully.", "I'll try a different approach.") OR an expression of uncertainty ("I'm not sure about this.", "I don't know how to solve this problem.").

Reply with exactly one letter: Y, N, C.

Problem: `{prompt}`

Prior steps:
`{prior_claims}`

New step: `{new_claim}`

E.4 Downstream evaluation prompts

E.4.1 FActScore: Likert scale

System prompt.

You are a careful, impartial judge rating short encyclopedia-style biographies. For each statement below, indicate your level of agreement on a 1-5 scale: 1 = strongly disagree, 2 = disagree, 3 = neither agree nor disagree, 4 = agree, 5 = strongly agree.

User prompt.

Rate the following biography of `{entity}` on three independent axes.

For each statement, return an integer from 1 (strongly disagree) to 5 (strongly agree):

1. **completeness**: "The biography covers the most important facts about `{entity}`, as one would expect to find in a short biography."
2. **fluency**: "The biography reads as coherent prose: sentences flow smoothly into one another, with no awkward phrasing, abrupt transitions, redundancy, or disjointed jumps between unrelated points."
3. **helpfulness**: "If a reader asked for a short biography of `{entity}`, this would be a satisfactory and helpful answer."

Biography:

```
"""
{text}
"""
```

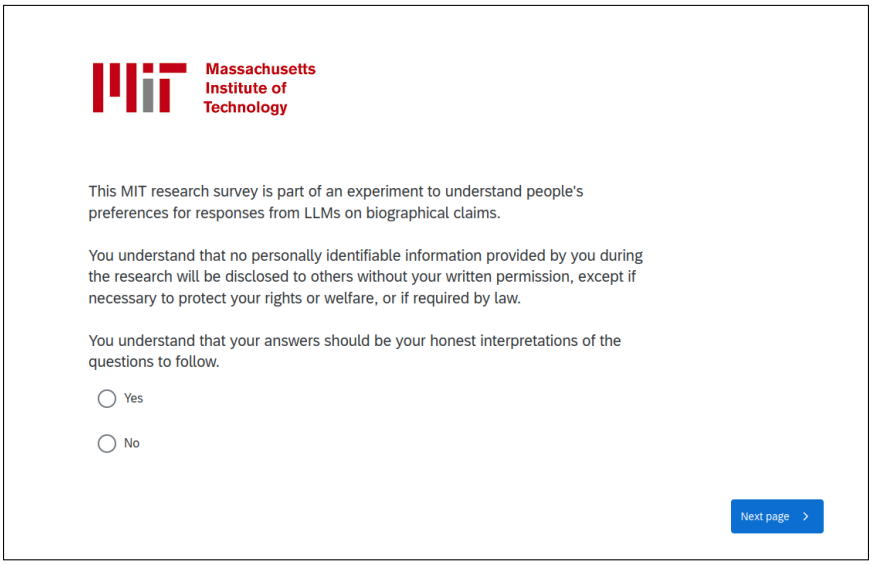
E.4.2 MATH: Completeness

We re-use the data labeling prompt from Appendix E.2. The downstream evaluation only consumes the `complete` field, discarding the rest of the output.

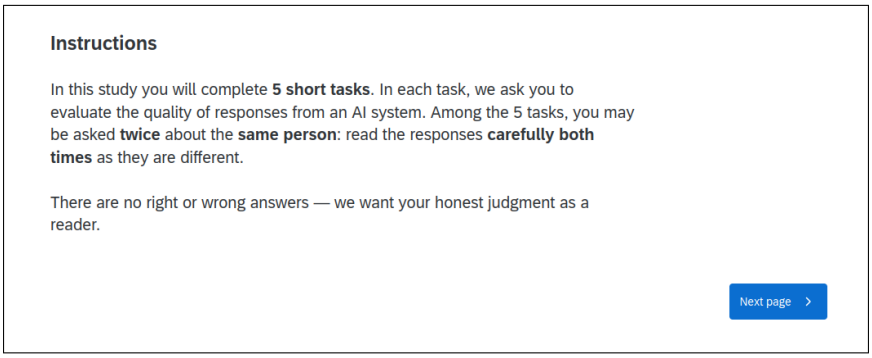
F Human-Study Details

We recruited annotators through the Prolific platform.[18] After giving informed consent, each annotator was shown a generated response and asked to complete the same Likert-style questionnaire as the automated judges (Section E.4.1); redacted screenshots are shown in Figure 17. In total, 200 annotators rated FActScore responses to 50 test prompts (coming from a single calibration/test split) at four target levels, with a median of 5 ratings per response. The study did not involve any personal data and was approved by an institutional review board. Annotators were compensated at an average rate of $44 per hour.

[18] https://www.prolific.com

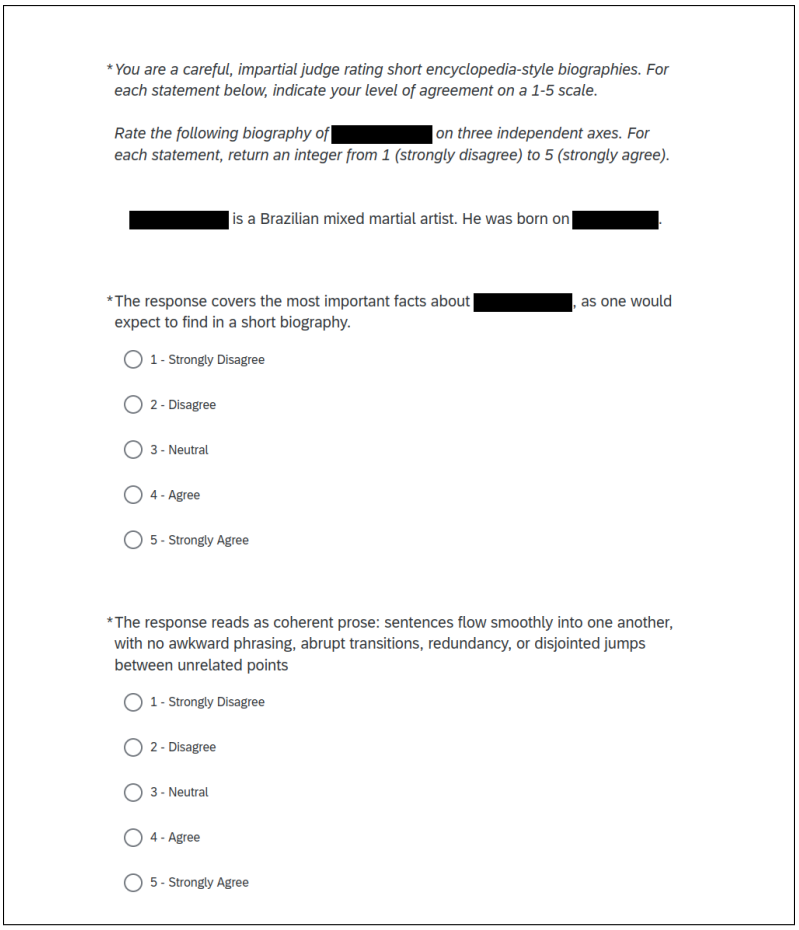


(a) Consent page.

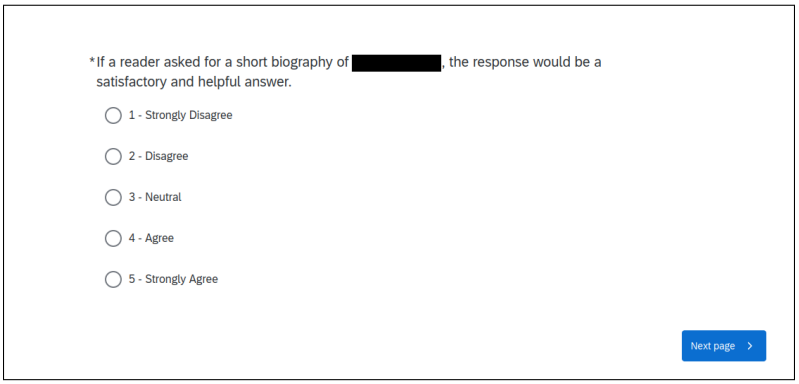


(b) Instructions page.

Figure 16. Prolific study: consent and instructions pages.



(a) Top of the page.



(b) Bottom of the same page.

Figure 17. Prolific study: a single rating task, split across two screenshots.

G Broader Impact

Our method aims to reduce hallucinations in LLM outputs and provide statistical guarantees on their factuality. This can lead to many positive societal effects, including more trustworthy deployment of such systems in high-stakes settings. However, since our guarantees are *marginal* rather than *conditional*, the system may be perceived as performing worse (or better) than expected from an individual user's point of view. Furthermore, since our method calibrates outputs with respect to an externally defined notion of factuality, it inherits the biases of whoever defines that notion; in the worst case, malicious or careless operation could thus effectively amount to censoring or suppressing outputs that disagree with the operator's worldview. As with any automated system, deployment of such a method to end-users thus requires careful design and monitoring to prevent misuse.

H Use of AI Assistance

We acknowledge the use of AI tools in the preparation of this manuscript. Specifically, we utilized language models (LMs) to assist with grammar and style editing, preparation of figures, general coding assistance, drafting and checking proofs, and other tasks. We have reviewed and edited the content to ensure accuracy and clarity, and we take full responsibility for the final version of the manuscript. We do not believe this constitutes important, original, or non-standard usage of such tools, but nonetheless include this disclosure for completeness.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state that we sample from a calibrated posterior with risk-control guarantees and improved downstream utility, which is supported by the theoretical results in Sections 2–3 and the experiments in Section 4.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We extensively discuss our method and the results we present, including limitations of our empirical study and of the core methodology, in Section 6. Furthermore, our empirical study in Section 4 highlights the different strengths and weaknesses of our approach compared to prior work.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Our theoretical results are stated alongside informal discussion of assumptions and proof strategies in Sections 2–3; complete, formal versions are given in Appendices A and B.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Section 4 specifies the datasets, models, hyperparameters, mixture-prior weights, sampling temperature, number of base generations, and split sizes; Appendix E lists all our prompts. All additional results in Appendix D that rely on alternative models or settings clearly list these changes.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Our code and data are available at github.com/theoxo/posterior-conformal-lms, alongside a README file specifying how to reproduce the main experimental results. The datasets (FActScore, MATH) and base models (Llama-3.1-8B-Instruct, Llama-3.3-70B-Instruct) are publicly available.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 4 reports the calibration/test split sizes (50 test, remainder calibration, averaged over 10 random splits), generation temperature (0.8), number of base samples (20), and mixture weights ($\beta = 0.05$). A more extensive ablation of the latter, which is the main hyperparameter of our method, is included in Appendix D.8.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

Justification: All plots in Section 4 report 95% confidence intervals (shaded regions) computed across 10 random calibration/test splits, as stated in the figure captions. We compute these intervals using the Student's t-distribution. Additionally, Appendices D.3, D.4 and also D.6 report bootstrapped 95% confidence intervals, and Figure 12 includes paired uncertainty intervals for the comparison to $M = 1$.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Section 4 reports that we serve the quantized Llama-3.3-70B-Instruct on a single NVIDIA RTX A6000 (48GB VRAM) at roughly 30s per prompt; scoring and labelling rely on OpenAI and Anthropic's public APIs and do not require any significant local computational resources.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics and believe that our work conforms to it.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.

- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Appendix G discusses both positive impacts (more trustworthy deployment via calibrated factuality control) and negative impacts (miscalibration as a false guarantee of truth, and potential misuse for censorship via biased scorers or calibration data).

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release any new models or data (other than experimental results with no risk for misuse).

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We cite FActScore [Min et al., 2023], MATH [Hendrycks et al., 2021], Llama-3.3-70B-Instruct [Grattafiori et al., 2024] and Llama-3.1-8B-Instruct [Grattafiori et al., 2024], all of which are well-established and standard assets in the field. Llama 3.3 is licensed under the Llama 3.3 Community License Agreement; Llama 3.1 is similarly licensed under the Llama 3.1 Community License Agreement. To the best of our knowledge, the datasets do not explicitly name their licenses.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: Our code and data released at https://github.com/theoxo/posterior-conformal-lms are accompanied by documentation in the form of a README and comments in the source code.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [Yes]

Justification: The human study (Appendix F) asked annotators the same Likert statements as the language model judges, whose full text is given in Appendix E.4.1; screenshots of the consent, instructions and rating pages are in Figures 16 and 17, and annotators were compensated at an average rate of $44 per hour.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: The human study (Appendix F) was approved by an institutional review board, and annotators gave informed consent before participating. The task, rating short biographies of public figures, posed minimal risk, and no personal data was collected.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: Appendix H declares the use of language models for grammar and style editing, figure preparation, general coding assistance, and proof drafting/checking. We do not believe this constitutes important, original, or non-standard usage of such tools, but nonetheless include this disclosure for completeness.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.