Conformal Steering of LLMs via Posterior Sampling

Nicolas Emmenegger [*1] **Theo X. Olausson** [*1] **Armando Solar-Lezama** [1] **Chara Podimata** [1]

Abstract

Large 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 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 an LLM posterior, 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). 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 & 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 n samples from the LM alongside their 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 altering 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 claims have strong interdependencies. In other words, while post-hoc filtering guarantees a bound on the expected number of responses containing a false claim, it often renders 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 & 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-)agressively filtered responses.

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*

[*]Equal contribution [1]Massachusetts Institute of Technology, Cambridge, MA, USA. Correspondence to: Nicolas Emmenegger <nemm@mit.edu>, Theo X. Olausson <theoxo@mit.edu>.

Accepted to the ICML 2026 Workshop on Statistical Frameworks for Uncertainty in Agentic Systems, Seoul, South Korea, 2026.
Copyright 2026 by the author(s).

*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 LM'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 from the LM, 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

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

1.1. Roadmap and Contributions

Modeling. In Sec. 2, we formulate conformal LM 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 LM distribution on high-confidence regions. While, as we show, this is cleanly motivated by a probabilistic formulation, our method's safety guarantees will be 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 hard 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 show experimentally 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 an LLM with vocabulary $\mathcal{V}$ and prompt x as a distribution $p_{\text{LLM}}(y \,|\, x) \in \Delta(\mathcal{V}^*)$ 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*

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

Our goal is to pick Φ so that substituting $p^{\Phi}_{\text{LLM}}(\cdot|x)$ 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 $p^{\Phi^\tau}_{\text{LLM}}(\cdot|x) = p^{\tau}_{\text{LLM}}(\cdot|x)$ defined via the parametrized potentials Φ^τ. The RHS of Equation (2) implicitly depends on the *normalization constant* $Z^\tau(x) = \int_{y \in \mathcal{V}^*} \Phi^\tau(y|x) \, \mathrm{d}p_{\text{LLM}}(y|x)$, 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.[1] 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

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

such that

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

where K is the (random) number of claims in the output. We will sometimes write $S(y \mid x) = \prod_{i\in[K]} s(c_i \mid x, \boldsymbol{c}_{<i})$ to denote completion-level scores. We build intuition for this definition by recalling the chain rule of probabilities. Assume 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, assume (see also a detailed derivation in Appendix A)that

$$\begin{aligned}&\mathbb{P}[\boldsymbol{c}_i \text{ is factual} \mid \boldsymbol{s}_{\leq i};\\ &\quad \text{previous claims } \boldsymbol{c}_{<i} \text{ are all factual}] = s_i.\end{aligned} \tag{4}$$

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 additional technicalities outlined in Appendix A, where the statement can be found in full detail, one can show the following observation.

Observation 1 (Idealized validity; informal)**.** Under an idealized, stepwise-calibrated[2] scoring function $s(\cdot)$, sampling $y \sim p_{\mathrm{LLM}}^{1-\alpha}(\cdot \mid x)$ induced by $\Phi^{1-\alpha}(y \mid x)$ satisfies $\mathbb{E}[\ell(x, y)] \leq \alpha$.

Intuitively, if the scores were stepwise-calibrated forecasts of conditional correctness given the preceding prefix, a martingale argument (Appendix A) would yield $\prod_{i\leq K} s_i = \mathbb{P}(y \text{ correct} \mid x, \boldsymbol{s}_{\leq K})$. 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.

2.3. 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 $Z_{\mathrm{LLM}}^\tau(x) = \int_{y\in\mathcal{V}^*} \Phi^\tau(y \mid x)\, \mathrm{d}p_{\mathrm{LLM}}(y \mid x) > 0$ 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 & 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_{LLM}^τ 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, $\beta > 0$ is a small pre-specified mass, and we require that the scorer assigns $S(\texttt{abstain} \mid x) = 1$ for all $x \in \mathcal{X}$. The resulting posterior then retains a factual *escape valve* at cutoff τ; abstaining is always factual. Analogously to (2), we define the mixture posterior $p_\beta^\tau(y \mid x) \propto p_\beta(y \mid x)\Phi^\tau(y \mid x)$. Conveniently for what follows, this also makes the conditional normalizing constant well behaved, i.e., $Z_\beta^\tau(x) = \int_{y\in\mathcal{V}^*} \Phi^\tau(y \mid x)\, \mathrm{d}p_\beta(y \mid x) \geq \beta$ for all τ. We often drop the subscript β and just write $p := p_\beta$ and $p^\tau = p_\beta^\tau$, and similarly $Z^\tau = Z_\beta^\tau$. Note that $p^0 = p_\beta^0 = p_\beta$ 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, yielding risk-controlling guarantees under very minimal assumptions, the scorer does have an impact on how stringently the system filters out generations (see Appendix D.3 for a scorer ablation). We follow prior work and elicit scores by asking an external LLM to rate its confidence in each claim on a scale from 0 to 1 (Mohri & Hashimoto, 2024). Empirically, this does not yield continuously-distributed scores, instead concentrating on a small, discrete set. Following Mohri & 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[3].

While having flexibility on the choice of scorer is conve-

[2] 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.

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

nient, 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 to consider in advance, and drawing a new calibration set for each (s, τ) pair — a prohibitively expensive process that would require 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

Having defined the potential family Φ^τ and the mixture prior p, we turn to the central question of the paper: how to pick τ so that posterior sampling controls risk at level α, using *only* samples from the base model. We will have access to a calibration set of prompts $\mathcal{C} = \{x_1, \ldots, x_n\}$ from P_X, and assume that $x_1, \ldots, x_n, x_{\text{test}}$ are exchangeable. We present our approach in 3 steps: (1) derivation of an idealized calibration rule under known normalization constant; (2) construction of a tractable empirical approximation; and (3) proof that this yields risk control for the resulting posterior sampler(s).

Known Normalization Constant. Assume that we have 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 was the case, then we could estimate the posterior risk using importance reweighting. To see this, note first that from Section 2.1: $\frac{p^\tau(y \mid x)}{p(y \mid x)} = \frac{\Phi^\tau(y \mid x)}{Z^\tau(x)}$. So, 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-weigh 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 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 gives an *idealized* calibration rule:

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

We briefly comment on this calibration rule. Note that since $\tau \mapsto Z^\tau(x_i)$ is a *monotone* non-increasing[4] map and by the thresholding nature of Φ^τ, each term in the sum is an upper bound on $\frac{\Phi^\tau(y_i \mid x_i)}{Z^\tau(x_i)}\ell(x_i, y_i)$. Fixing the denominator to not depend on τ has an additional crucial property: the ratio is now monotone in τ. Therefore, the LHS of the constraint in (6) inherits this *monotonicity*, yielding our first risk control guarantee.

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

Unknown $Z^\tau(x)$. In practice, computing $Z^\tau(x)$ is intractable[5], and the rule (6) cannot be evaluated. A natural idea is to 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 this end, we draw M i.i.d. responses $y_{ij} \sim p_{\text{LLM}}(\cdot \mid x_i)$ (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

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

The explicit abstention path implies $\hat{Z}^\tau(x_i, T_i) \geq \beta$ which makes $\hat{p}^\tau$ well-defined for all $\tau \in [0, 1]$.

Calibrating against the empirical posterior. The definition of $\hat{p}$ signifies that we are essentially stratifying the sampling from both modes of the mixture prior. This will naturally lead to a calibration routine that does *not* control the risk under the population posterior p^τ, but instead under the empirical posterior $\hat{p}^\tau$. To this end, given T_i, we consider the conditional risk under $\hat{p}^\tau$, namely

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

[5] 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 $x_1, \ldots, x_n$; particle budget M; abstention mass β; target level α

for $i = 1, \ldots, n$ **do**
  for $j = 1, \ldots, M$ **do**
    Sample $y_{ij} \sim p_{\text{LLM}}(\cdot \mid x_i)$, score it to obtain $\Phi^\tau(y_{ij} \mid x_i)$, and 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 $\hat{\tau} = \inf\left\{\tau \,\middle|\, \frac{1}{n+1}\sum_{i=1}^{n} \bar{H}_i(\tau) + \frac{1}{n+1} \le \alpha\right\}$.

$\mathbb{E}_{y \sim \hat{p}^\tau(\cdot \mid x_i, T_i)}[\ell(x_i, y)]$, and write the shorthand

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

because the abstention atom contributes $\ell(x, \texttt{abstain}) = 0$ to the numerator and can be omitted. $\hat{H}_i(\tau)$ 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[6], $\hat{H}_i$ need not be monotone: pruning a correct particle at a slightly higher threshold while leaving incorrect particles to survive can transiently raise the average loss among the accepted particles. We restore monotonicity sample-pathwise via the monotone upper envelope:

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

a non-increasing pointwise upper bound on $\hat{H}_i$. Monotonization is the lever that makes conformal-risk-control style arguments work. Empirically (see Sec. 4), $\hat{H}_i$ is close to monotone for moderate M, and $\bar{H}_i$ introduces little slack.

The full procedure is given in Algorithm 1. We can give the following conformal guarantee, which 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 $x_1, \ldots, x_n, x_{\text{test}}$ be exchangeable, and assume that conditionally, all calibration and test particles are drawn independently with $y_{ij} \sim p_{\text{LLM}}(\cdot|x_i)$ and $(y_{\text{test}})_j \sim p_{\text{LLM}}(\cdot|x_{\text{test}})$. Let $\hat{\tau}$ be the output of Algorithm 1. If $y_{\text{test}} \sim \hat{p}^{\hat{\tau}}(\cdot|x_{\text{test}}, T_{\text{test}})$, then $\mathbb{E}[\ell(x_{\text{test}}, y_{\text{test}})] \le \alpha$, 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)**.** *Fix abstention mass $\beta > 0$, and let $T_M = (y_1, \ldots, y_M)$ with $y_j \overset{i.i.d.}{\sim} p_{\text{LLM}}(\cdot \mid x)$. Then for P_X-a.e. x,*

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

Note that this is an almost surely statement about 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.

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

In other words, the calibration guarantee transfers asymptotically to any deployment sampler that targets the population posterior. This means that we achieved the goal we set out to, modulo error that shrinks as M grows. We also point out that we need not necessarily enter the asymptotic regime for the finite-particle, stratified way of constructing the empirical posterior to make sense. Different ways to sample from the posterior are possible that still exactly or approximately inherit the finite sample guarantee in Theorem 3.2. For instance, Sequential Monte Carlo steering (Lew et al., 2023; Zhao et al., 2024; Loula et al., 2025) fixes the initial number of particles, in which case our method amounts to extending the proposal distribution with an appropriately weighted abstention particle; see Sections 5 and 6 for further discussion on efficient sampling and particle filtering methods.

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

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 the standard test set of 500 problem instances from MATH (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`[7] (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; App. D.3 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 $\beta = 0.1$ 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.4.

Baseline. We compare against Mohri & Hashimoto (2024), the closest prior work. We re-implement Mohri & 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`.

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

4.1. Validity

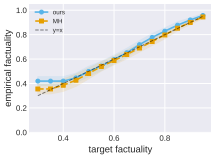


Figure 2. Empirical vs. target factuality on FActScore; ours (—) and `MH` (—). Both attain valid coverage.

Figure 2 shows the empirical risk obtained versus the desired target $1 - \alpha$ on FActScore (results on MATH in App. 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 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.

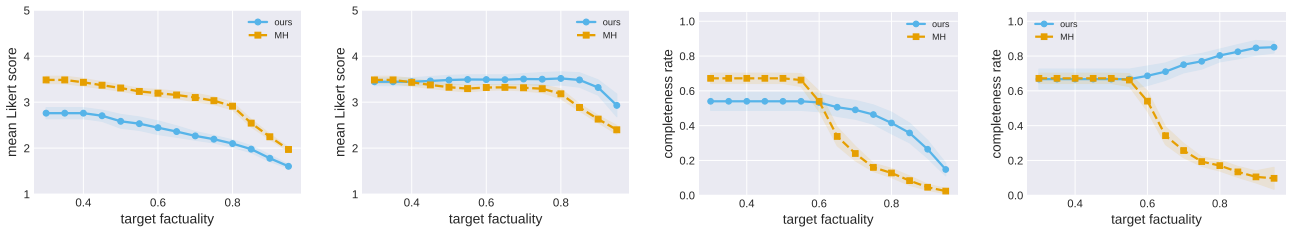
4.2. Utility

We use LLM-judge scoring 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). Abstentions are naturally scored as incomplete. To reduce annotation cost, we subsample a single generation[8] 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 1 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

[8]For our method, from the posterior. For `MH`, from the base model with subsequent filtering.



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

Figure 1. 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.

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*, our method underperforms MH. This indicates that the LM 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 targets ≥ 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.

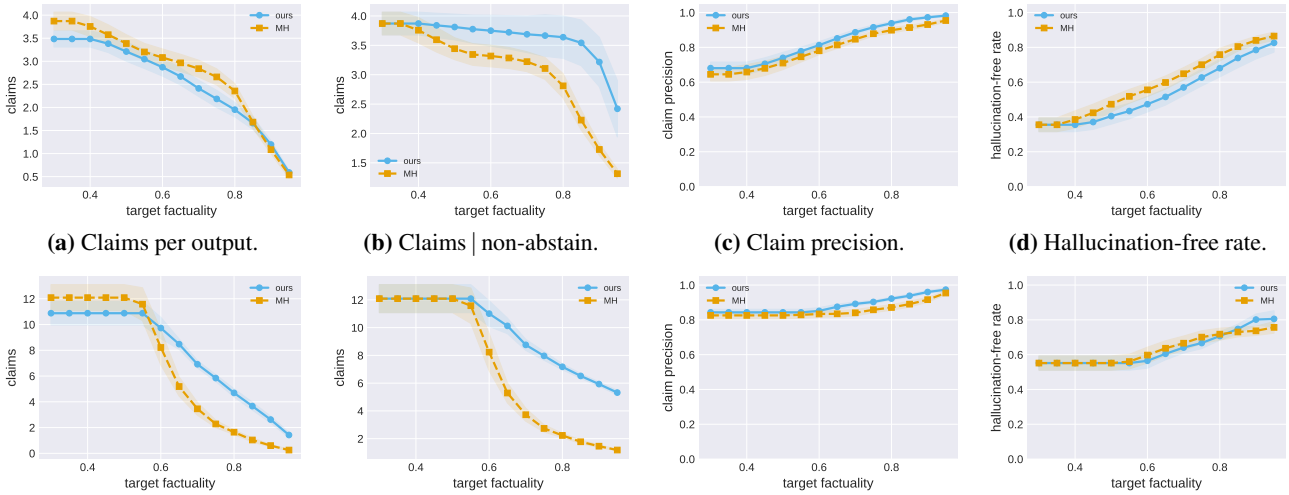
Proxy results. Figure 3 shows the claim-level proxy metrics, with FActScore on the top row and MATH on the bottom. *First column*: At intermediate targets, our method emits roughly as many claims per instance as MH. The mixture prior places β of the mass on zero-claim abstentions, deflating our method's average at low targets; this reverses at high targets, where MH increasingly produces 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 LM). 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., 2025; 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 & 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. Both the goal and the approach presented in this work are thus orthogonal from such efforts.

Particle methods for semantic steering of language models. Our method is inspired by a line of work that frames semantically constrained generation as posterior sampling (Lew et al., 2023; Zhao et al., 2024; Loula et al., 2025). Similarly to our setup, these works treat a language model as a probabilistic prior and then build a suitably defined posterior distribution. However, their goal is different than ours, predominantly being concerned with sampling efficiency. Meanwhile, we focus on how to leverage *unreliable* potentials, a challenge which has not been explored by this



(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).

literature.

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 (c.f. App. C). We refer to the works of Fannjiang et al. (2022) and Prinster et al. (2024) for a more thorough treatment of conformal guarantees under more general (covariate) shifts.

6. Discussion: Limitations & Future Directions

The β parameter. This parameter governs the mixture prior between the LM completions 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 can reduce the informativeness of non-abstaining outputs. Smaller β values preserve more mass on LM completions, improving utility when the model is capable, but leave less fallback mass when high-confidence completions are rare. Additionally, the choice of β and its relative scale compared to M 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).

Measuring downstream utility. Our empirical evaluation measures downstream utility using LLM-judge scoring, focusing on properties such as completeness, fluency, and mathematical coherence. While this provides a scalable first validation, it is not a substitute for human evaluation. A natural next step is to complement these experiments with human annotations. Moreover, human preference annotations would shed more light on the trade-off between the utility of responses vs abstentions, as discussed in Section 4. Such studies would thus help clarify whether the utility improvements measured by LLM judges correspond to user-perceived improvements in realistic deployment settings.

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 LM, 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.

References

Angelopoulos, A., Bates, S., Fisch, A., Lei, L., and Schuster, T. Conformal risk control. In *International conference on learning representations*, volume 2024, pp. 55198–55218, 2024a.

Angelopoulos, A. N., Barber, R. F., and Bates, S. Theoretical foundations of conformal prediction. *arXiv preprint arXiv:2411.11824*, 2024b.

Angelopoulos, A. N., Bates, S., Candès, E. J., Jordan, M. I., and Lei, L. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19(2):1641–1662, 2025.

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

Damani, M., Puri, I., Slocum, S., Shenfeld, I., Choshen, L., Kim, Y., and Andreas, J. Beyond binary rewards: Training lms to reason about their uncertainty. *International conference on learning representations*, 2026.

Dehghani, F., Dehghani, R., Naderzadeh Ardebili, Y., and Rahnamayan, S. Large language models in legal systems: A survey. *Humanities and Social Sciences Communications*, 12(1):1977, 2025.

Detommaso, G., Bertran, M., Fogliato, R., and Roth, A. Multicalibration for confidence scoring in llms. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Fan, D., Delsad, S., Flammarion, N., and Andriushchenko, M. Halluhard: A hard multi-turn hallucination benchmark. *arXiv preprint arXiv:2602.01031*, 2026.

Fannjiang, C., Bates, S., Angelopoulos, A. N., Listgarten, J., and Jordan, M. I. Conformal prediction under feedback covariate shift for biomolecular design. *Proceedings of the National Academy of Sciences*, 119(43):e2204569119, 2022.

Gaber, F., Shaik, M., Allega, F., Bilecz, A. J., Busch, F., Goon, K., Franke, V., and Akalin, A. Evaluating large language model workflows in clinical decision support for triage and referral and diagnosis. *npj Digital Medicine*, 8 (1):263, 2025.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Hébert-Johnson, U., Kim, M., Reingold, O., and Rothblum, G. Multicalibration: Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning*, pp. 1939–1948. PMLR, 2018.

Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D., and Steinhardt, J. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.

Lew, A. K., Zhi-Xuan, T., Grand, G., and Mansinghka, V. K. Sequential monte carlo steering of large language models using probabilistic programs. *arXiv preprint arXiv:2306.03081*, 2023.

Loula, J., LeBrun, B., Du, L., Lipkin, B., Pasti, C., Grand, G., Liu, T., Emara, Y., Freedman, M., Eisner, J., et al. Syntactic and semantic control of large language models via sequential monte carlo. In *International Conference on Learning Representations*, volume 2025, pp. 64758–64791, 2025.

Min, S., Krishna, K., Lyu, X., Lewis, M., Yih, W.-t., Koh, P., Iyyer, M., Zettlemoyer, L., and Hajishirzi, H. 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*, pp. 12076–12100, 2023.

Mohri, C. and Hashimoto, T. Language models with conformal factuality guarantees. In *International Conference on Machine Learning*, pp. 36029–36047. PMLR, 2024.

Prinster, D., Stanton, S. D., Liu, A., and Saria, S. Conformal validity guarantees exist for any data distribution (and how to find them). In *Proceedings of the 41st International Conference on Machine Learning*, pp. 41086–41118, 2024.

Quach, V., Fisch, A., Schuster, T., Yala, A., Sohn, J. H., Jaakkola, T., and Barzilay, R. Conformal language modeling. In *International Conference on Learning Representations*, volume 2024, pp. 11654–11681, 2024.

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

Rubin-Toles, M., Gambhir, M., Ramji, K., Roth, A., and Goel, S. Conformal language model reasoning with coherent factuality. In *The Thirteenth International Conference on Learning Representations*, 2025.

Sadhuka, S., Prinster, D., Fannjiang, C., Scalia, G., Regev, A., and Wang, H. E-valuator: Reliable agent verifiers with sequential hypothesis testing. *arXiv preprint arXiv:2512.03109*, 2025.

Saunders, C., Gammerman, A., and Vovk, V. Transduction with confidence and credibility. In *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*, 1999.

Tibshirani, R. J., Foygel Barber, R., Candes, E., and Ramdas, A. Conformal prediction under covariate shift. *Advances in neural information processing systems*, 32, 2019.

Van Calster, B., Nieboer, D., Vergouwe, Y., De Cock, B., Pencina, M. J., and Steyerberg, E. W. A calibration hierarchy for risk models was defined: from utopia to

empirical data. *Journal of clinical epidemiology*, 74: 167–176, 2016.

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

Wu, M., Zhou, C., Bates, S., and Jaakkola, T. Thought calibration: Efficient and confident test-time scaling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 14302–14316, 2025.

Zhao, S., Brekelmans, R., Makhzani, A., and Grosse, R. B. Probabilistic inference in language models via twisted sequential monte carlo. In *International Conference on Machine Learning*, pp. 60704–60748. PMLR, 2024.

Zhou, C., Wang, Z., Wu, M., Zhu, Q. J., Shi, F. C., Wang, C., Wilson, A., Jaakkola, T., and Bates, S. 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 $k > K$, 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[9]. 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 $y \sim p_{\text{LLM}}(\cdot \mid x)$, 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 $\tau \in [0, 1]$,*

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

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.

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

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

We briefly comment on the steps:

(i) follows because
$$\mathbb{E}[Z_k \mid \mathcal{F}_k, Z_{<k}] = \mathbf{1}\{A\}\,\mathbb{E}[Z_k \mid \mathcal{F}_k, A] + \mathbf{1}\{A^c\}\,\mathbb{E}[Z_k \mid \mathcal{F}_k, A^c]$$
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[10], 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
$$\mathbb{E}\left[\ell(x,y) \mid \prod_{i=1}^{K} s_i \geq \tau\right] = \mathbb{P}\left(\ell(x,y) = 1 \mid \prod_{i=1}^{K} s_i \geq \tau\right) = \mathbb{P}\left(Z_{\leq K} = 0 \mid \prod_{i=1}^{K} s_i \geq \tau\right).$$
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
$$\begin{aligned}
\mathbb{P}(Z_{\leq K} = 1 \mid \textstyle\prod_{i=1}^{K} s_i \geq \tau) &= \frac{\mathbb{P}(Z_{\leq K} = 1 \text{ and } \prod_{i=1}^{K} s_i \geq \tau)}{\mathbb{P}(\prod_{i=1}^{K} s_i \geq \tau)} \\
&= \frac{\mathbb{E}\left[\mathbf{1}\left\{Z_{\leq K} = 1 \text{ and } \prod_{i=1}^{K} s_i \geq \tau\right\}\right]}{\mathbb{P}(\prod_{i=1}^{K} s_i \geq \tau)} \\
&= \frac{\mathbb{E}\left[Z_{\leq K}\mathbf{1}\left\{\prod_{i=1}^{K} s_i \geq \tau\right\}\right]}{\mathbb{P}(\prod_{i=1}^{K} s_i \geq \tau)}.
\end{aligned}$$

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

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

[10]The maximum length of the LLM response

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

Proof. (1): $\bar{w}_i(\tau)\ell_i = \frac{\ell_i}{Z^{S(y_i\mid x_i)}(x_i)}\mathbf{1}\{S(y_i\mid x_i)\geq\tau\}$ is constant for $\tau \leq S(y_i\mid x_i)$ and zero for $\tau > S(y_i\mid x_i)$. (2): For $\tau \leq S(y_i\mid x_i)$, $Z^{S(y_i\mid x_i)}(x_i) \leq Z^\tau(x_i)$ (since Z^τ is non-increasing), so $\bar{w}_i(\tau) \geq w_i(\tau)$; for $\tau > S(y_i\mid x_i)$, 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 $\ell, \Phi^\tau \in \{0,1\}$. □

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:

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

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

Proof. For any τ in the feasible set of $\hat{\tau}_{\text{id}}$, 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 $\tau^*_{\text{id}} \leq \hat{\tau}_{\text{id}}$ 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[11], 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

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

□

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.

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

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

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 $[0,1]$. □

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*

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

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 & 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.[12]

Lemma B.6. *Assume the distribution of the scores is continuous*[13]*. Let $h : \mathcal{V}^* \times \mathcal{S} \to \{0,1\}$ be a fixed measurable map. Define the function class*

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

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

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

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

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

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

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

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

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

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

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

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

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

□

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

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

We are ready to prove Lemma 3.3.

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

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

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

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 $\{y \mapsto 1\{S(y \mid x) \geq \tau\} : \tau \in [0, 1]\}$ and $\{y \mapsto \ell(x, y) 1\{S(y \mid x) \geq \tau\} : \tau \in [0, 1]\}$ have finite Bracketing number, by Lemma B.6. By Theorem 2.4.1. in (van der Vaart & 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

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

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

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

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

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

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}:

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

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,

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

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

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

and hence

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

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-reweights 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 $E = \{U_1, \ldots, U_n, U_{n+1}\}$

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

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 $y_{n+1} \sim p^\tau(\cdot \mid x_{n+1})$ for *fixed* τ, we'd have

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

In the known Z^τ case, this could suggest[14] an oracle selection rule

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

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 hallucinated calibration set 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, the expected denominator in τ^*_{SN} coincides with that in τ^*_{id}. Writing $w_i(\tau) = \Phi^\tau(y_i \mid x_i)/Z^\tau(x_i)$, a direct calculation using the tower rule gives

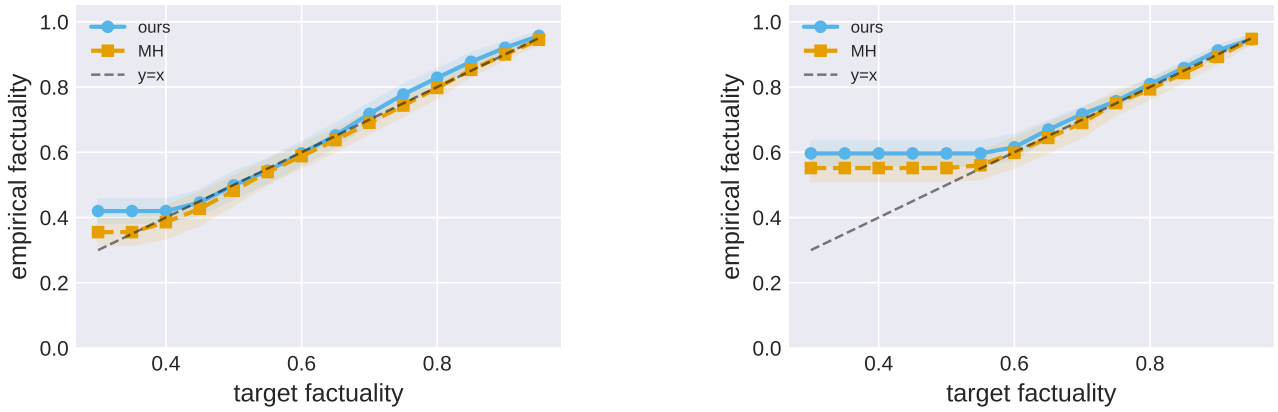
$$\mathbb{E}\left[\sum_{i=1}^{n+1} w_i(\tau)\right] = \sum_{i=1}^{n+1} \mathbb{E}\left[\mathbb{E}\left[\frac{\Phi^\tau(Y_i \mid X_i)}{Z^\tau(X_i)} \,\middle|\, X_i\right]\right] = \sum_{i=1}^{n+1} \mathbb{E}\left[\frac{Z^\tau(X_i)}{Z^\tau(X_i)}\right] = n+1.$$

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. Extensive Experimental Results

D.1. MATH Calibration Validity

Figure 4 shows the results of the empirical validity experiment (Figure 2) 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.



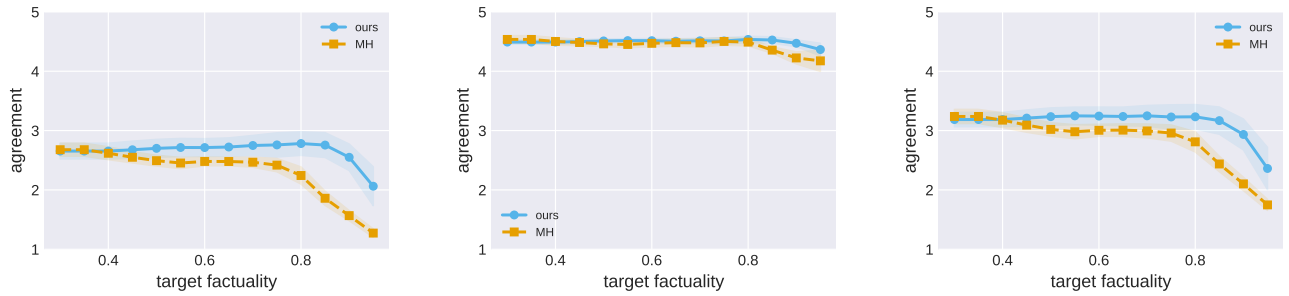
(a) FActScore. **(b)** MATH.

Figure 4. Target vs. observed factuality on FActScore (a) and MATH (b); ours (blue) and MH (orange). 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 1 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 1 is the arithmetic mean of these axes.

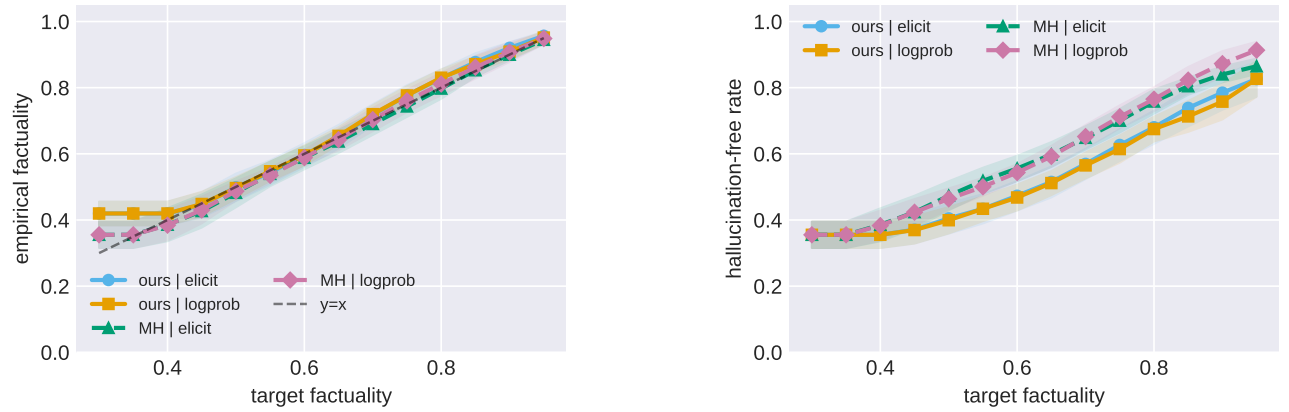
[14] Ignoring for a second that monotonizing this objective is also trickier than if the denominator is fixed.



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

Figure 5. Breakdown of the FActScore Likert results from Figure 1 (panel b) per statement that the judge is asked to rate its agreement with: (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. Alternative scoring functions and base models



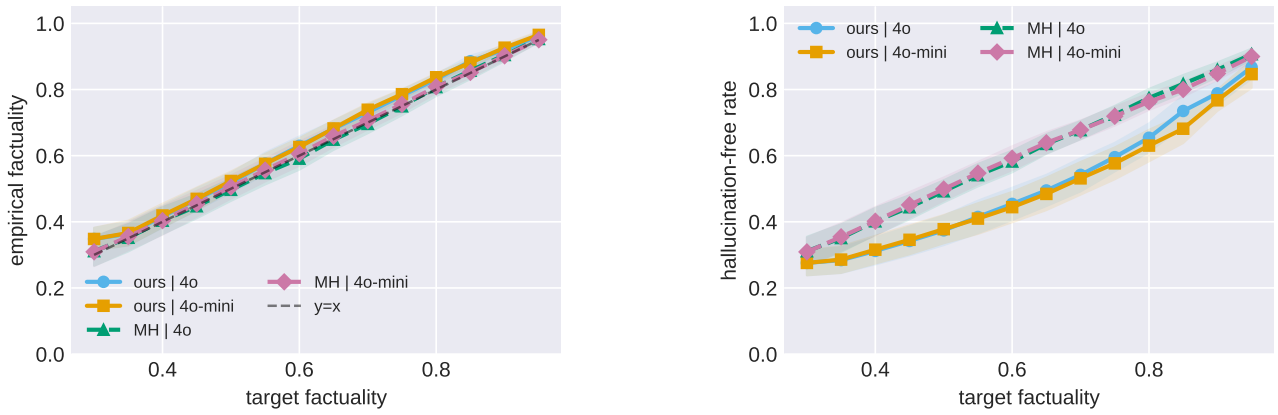
(a) Target vs. observed factuality. **(b)** Strict factuality of non-abstaining outputs.

Figure 6. Scorer-elicitation ablation on FActScore: confidence-rating elicitation (main text) vs. raw token log-probabilities. 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` judgement; we compute this scorer as $s = (p_{\mathtt{T}} + p_{\mathtt{N}})/(p_{\mathtt{T}} + p_{\mathtt{N}} + p_{\mathtt{F}})$ 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. 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. Overall, however, these differences are small, indicating that calibrated factuality successfully translates to downstream utility.



(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 scoring, all on Llama-3.1-8B-Instruct generations (in contrast to the larger Llama-3.3-70B-Instruct used elsewhere). Averaged over 10 splits; shaded area shows 95% CI.

D.4. Ablating β at $M = 20$ with fixed-budget rejection sampling

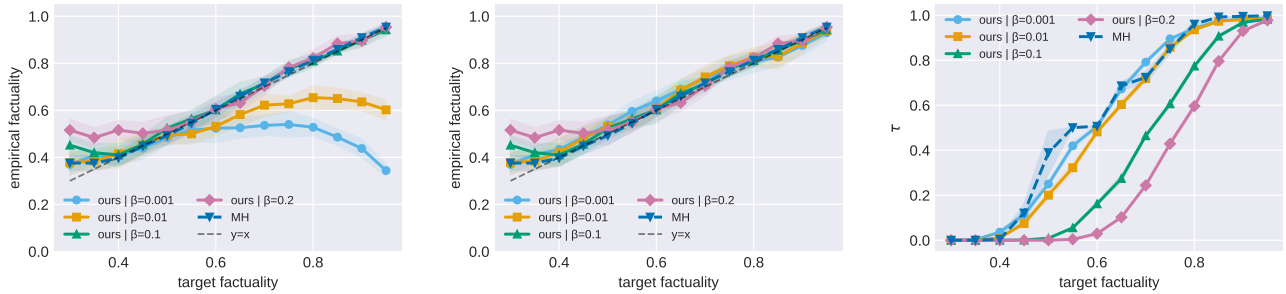
This appendix probes how our finite-particle guarantee for rejection sampling (Lemma 3.3 and Corollary 3.4) interacts with β and a fixed-budget rejection sampling deployment procedure. We fix $M = 20$ as in the main experiments, but at deployment we run rejection sampling up to a hard cap of 50 attempts; if a non-abstaining, non-accepted generation has not been reached 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 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.

Calibration validity. Figures 8a and 8b 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 $\beta = 0.1$ and $\beta = 0.2$, coverage is empirically valid and tight. 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 9 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 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

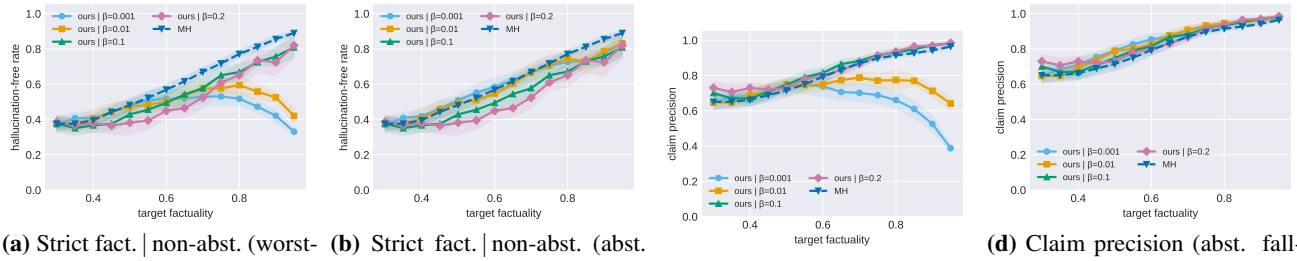


(a) Target vs. observed factuality (worst-case fallback).

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

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

Figure 8. 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). Averaged over 10 splits; shaded area shows 95% CI.



(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 9. Proxy metrics under the fixed-budget rejection sampling procedure on FActScore, comparing the worst-case (fully-false) and `abstention` budget-exhausted fallbacks.

because a lower forced probability of abstention in the mixture prior yields higher values of $\hat{\tau}$ (Figure 8c), thus applying stricter checks on the outputs.

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 prompt used for the alternative logprob-based score function used in Appendix D.3.

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:

MATH (logprob elicitation).

> 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.