Forecasting Model Success at Inference Time: Calibrated Probabilistic Forecasts for Cost-Optimal LLM Cascades

Anonymous Authors[1]

Abstract

Forecasting is often framed as predicting external future events. We study a shorter-horizon but operationally central forecast inside AI systems: will a cheaper language model succeed on this query, or should the system pay for a stronger model? We instantiate this view as UCCI, an uncertainty-calibrated cascaded inference system. UCCI (i) extracts a token-margin signal from the small model's generation, (ii) calibrates it with isotonic regression into a probability forecast of small-model error, and (iii) escalates to the large model only when the forecasted accuracy gain justifies the marginal cost. We prove that, under calibrated forecasts and a constant-gain escalation model, threshold policies are cost-optimal, and we state the calibration-data rate that controls expected calibration error. On a 75,000-query production named entity recognition workload, calibration reduces ECE from 0.12 to 0.03, and the resulting cascade reaches micro-F1 of 0.91 at 31% lower cost than always using the large model (95% CI [27%, 35%]). The same result also identifies the failure mode: under margin pathologies or distribution shift, the forecast loses calibration and the routing threshold loses its decision-theoretic meaning. The paper's central claim is that forecast quality, not raw model confidence, is the load-bearing object in cost-aware AI inference.

1. Introduction

Probabilistic forecasting has a simple contract: a system reports a probability, reality reveals an outcome, and a proper scoring rule evaluates whether the probability was honest (Gneiting & Raftery, 2007). That contract appears inside deployed AI systems whenever a controller must act before it observes whether a model will be correct. Cascaded inference is a clean example. A small model has already produced an answer; a large model could be called at higher cost; the router must decide whether the small answer is likely to be good enough.

This is a forecasting problem. The event is internal rather than external: "the small model will be wrong." The time horizon is milliseconds rather than weeks. But the statistical requirements are the same. A routing score used as a probability should be calibrated, and a routing rule should be evaluated not only by accuracy but by the cost of acting on the forecast. This perspective is especially relevant for AI forecasting: if general AI systems are expected to make reliable probabilistic judgments about the world, they should first make calibrated judgments about their own near-term success.

The forecasting reframe. Existing cost-aware LLM routers often threshold raw confidence signals such as entropy, maximum probability, or top-token margin (Chen et al., 2024a; Schuster et al., 2022). Those signals may rank examples usefully, but they are not probabilities. Modern neural networks are often miscalibrated (Guo et al., 2017; Ovadia et al., 2019); treating an uncalibrated score as a probability biases every downstream cost decision. UCCI instead makes the router an explicit forecaster: it predicts $\Pr[\text{small model wrong} \mid \text{small-model signal}]$, checks reliability, and chooses the cost-minimizing action under an accuracy constraint.

Contributions. We make three claims.

- **A calibrated inference-time forecast.** We aggregate per-token top-1 minus top-2 margins from the small model and calibrate the scalar score with isotonic regression (Zadrozny & Elkan, 2002), producing $\hat{p}(x) \approx \Pr[\text{small model wrong} \mid x]$.
- **A decision-theoretic router.** Under calibrated forecasts and a constant-gain escalation model, the cost-optimal cascade is a threshold policy on $\hat{p}(x)$. The threshold is the dual price of accuracy.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

- **An empirical forecast-quality result.** On 75,000 labeled NER queries, calibration drops ECE from 0.12 to 0.03, and UCCI reaches F1=0.91 at 31% lower cost than a large-only system, outperforming entropy, conformal, and FrugalGPT-style routers at matched accuracy.

The takeaway is that cascaded inference is more than an efficiency trick. It is a compact testbed for AI forecasting: the decision improves only when the probability forecast is reliable.

2. Problem Setup

Let $x \sim \mathcal{D}$ be a query with ground truth y. A small model f_s and large model f_ℓ have per-query costs $c_s < c_\ell$. A routing policy $\pi : \mathcal{X} \to \{s, \ell\}$ returns prediction $\hat{y}_\pi(x)$ and incurs cost $C_\pi(x)$. Given a target accuracy τ, we seek

$$\min_\pi \mathbb{E}[C_\pi(x)] \quad \text{s.t.} \quad \mathbb{E}[\mathrm{Acc}(\hat{y}_\pi(x), y)] \geq \tau. \tag{1}$$

Routing as one-step forecasting. Define the binary event $e(x) = \mathbf{1}\{\hat{y}_s \neq y\}$. The router's job is to forecast $\Pr[e(x) = 1]$ using information available after the small model runs but before the large model is called.

Token-margin signal. For a generation of length T with top-1 and top-2 next-token probabilities $p_{t,1}, p_{t,2}$, let $m_t = p_{t,1} - p_{t,2}$ and

$$u(x) = 1 - \frac{1}{T} \sum_{t=1}^{T} m_t. \tag{2}$$

Greedy decoding makes $u(x)$ well-defined.

Calibration. We learn an isotonic map $g : [0, 1] \to [0, 1]$ on a held-out calibration set $\{(u_i, e_i)\}_{i=1}^n$, yielding $\hat{p}(x) = g(u(x)) \approx \Pr[e(x) = 1 \mid u(x)]$. We evaluate the forecast with reliability diagrams and expected calibration error (ECE), the same diagnostics used for probabilistic forecasts of real-world events.

Threshold policy. Given $\hat{p}(x)$, define

$$\pi_\theta(x) = \begin{cases} s & \hat{p}(x) \leq \theta, \\ \ell & \hat{p}(x) > \theta. \end{cases} \tag{3}$$

On a validation set where both models have been run, we choose θ as the lowest-cost threshold satisfying the target accuracy.

3. Theoretical Analysis

Theorem 3.1 (Optimality of calibrated threshold routing)**.** *Assume $c_\ell > c_s$, $\hat{p}(x) = \Pr[e(x) = 1 \mid u(x)]$, and the expected accuracy gain from escalating is $\Delta(u) = a_\ell - a_s(u)$ with a_ℓ constant and $a_s(u) = 1 - \hat{p}(u)$. Among policies that depend only on $u(x)$, a cost-minimizing policy subject to Equation (1) is a threshold policy on $\hat{p}(x)$.*

Proof sketch. Escalating a query costs $c_\ell - c_s$ and changes expected accuracy by $\Delta(u) = a_\ell - 1 + \hat{p}(u)$, which is monotone in $\hat{p}(u)$. The constrained optimum buys accuracy in decreasing order of accuracy gain per unit cost, which is equivalent to escalating examples in decreasing $\hat{p}(u)$ until the accuracy constraint binds. This ordering is implemented by a threshold on $\hat{p}$. Full proof in Section A. □

Proposition 3.2 (Calibration rate)**.** *Under standard isotonic-regression regularity assumptions, isotonic calibration on n examples satisfies $\mathbb{E}\|\hat{g} - g^\star\|_2^2 = O(n^{-2/3})$ (Barlow et al., 1972), hence $\mathrm{ECE} = O(n^{-1/3})$ up to empirical concentration terms.*

Forecasting connection. The threshold is the cascade analogue of a Neyman-Pearson cutoff: it is the optimal cutoff for a binary forecast over an event that has a real decision cost. If calibration assumption fails, the threshold no longer corresponds to a uniform price on errors, and the policy is generically suboptimal. That makes calibration a governing assumption, not a post-processing detail.

4. Experiments

4.1. Setup

Task. NER on natural-language search queries from a production photo-management system. Outputs are JSON objects with up to 6 entity fields: camera, lens, aperture, shutter speed, ISO, and focal length. The dataset contains 75,000 manually labeled queries; 89% contain at least one entity; mean length is 6.2 words. Splits are 22,500 calibration, 15,000 validation, and 37,500 test queries.

Models and cost. f_s is a 4B instruction-tuned LLM and f_ℓ is a 12B instruction-tuned LLM, served with the same prompt, greedy decoding, max tokens 256, and vLLM (Kwon et al., 2023) on H100 GPUs. Mean end-to-end latency over 100 queries is 47.2 ± 3.1 ms for the small model and 142.3 ± 8.7 ms for the large model, giving normalized costs $c_s = 1.00$ and $c_\ell = 3.02$.

Evaluation. We select θ on validation and report test micro-F1 and measured normalized cost. Baselines are always-small, always-large, entropy thresholding, conformal routing (Angelopoulos & Bates, 2021; Sadinle et al., 2019), and FrugalGPT-style confidence routing (Chen et al., 2024a).

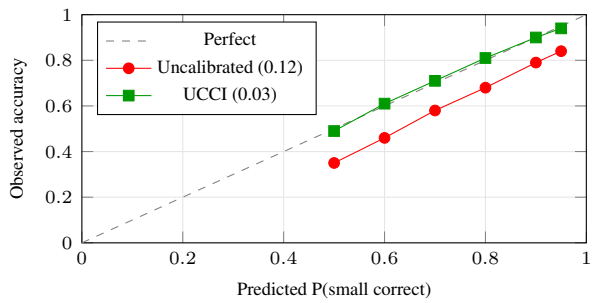

Figure 1. Reliability diagram. Calibration turns the small model's token-margin signal into a usable probability forecast.

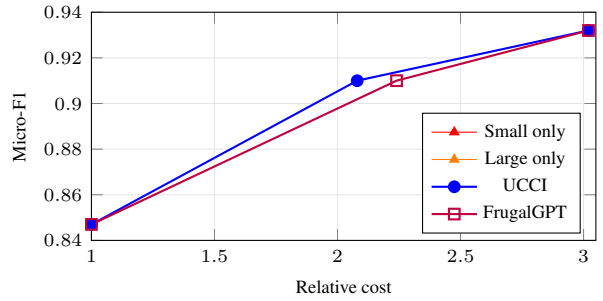

Figure 2. Cost-accuracy frontier on the test set. UCCI reaches the target F1 at lower measured cost than uncalibrated routing.

4.2. Results

Single-model anchors. On the test set, the small model reaches micro-F1 0.847 at cost 1.00, while the large model reaches micro-F1 0.932 at cost 3.02.

Forecast quality. Isotonic calibration on 22,500 examples reduces ECE from 0.12 for the raw margin signal to 0.03 for UCCI (95% CI [0.02, 0.04] via bootstrap). Figure 1 shows the reliability diagram. The uncalibrated signal is overconfident at high confidence, while the calibrated forecast tracks the diagonal.

Cost-accuracy frontier. Figure 2 traces the cost-accuracy frontier. At F1=0.91, UCCI reaches cost 2.08 versus 3.02 for large-only, a 31% reduction (95% CI [27%, 35%]). Table 1 compares routing methods at the matched-accuracy operating point.

Assumption check. The theorem assumes approximately constant large-model accuracy on escalated examples. At the F1=0.91 operating point, the large model's F1 on the escalated subset is 0.928 versus 0.932 overall, a difference of 0.004 within bootstrap noise. This supports the threshold model in our regime, but it should be rechecked under shift.

4.3. When the forecast fails

The forecasting view makes failure modes legible. Calibration can fail under heavy-tailed margin distributions, where a small subpopulation near zero margin dominates calibration error, or under distribution shift, where the calibration distribution no longer matches deployment. In either case, $\hat{p}$ stops being a trustworthy forecast and threshold routing no longer has a decision-theoretic interpretation. The right diagnostic is the same one used in forecasting: reliability under the deployment distribution.

Table 1. Routing methods at the F1=0.91 operating point.

Method	Cost	ΔF1 vs. target
UCCI	**2.08**	0.00
Conformal	2.18	0.00
FrugalGPT-style	2.24	0.00
Entropy	2.31	0.00
Large only	3.02	+0.02

4.4. Ablations

Isotonic calibration (ECE 0.03) beats temperature scaling (ECE 0.08) and uncalibrated routing on the same signal. Token margin beats predictive entropy and maximum probability at matched calibration data. Per-entity heterogeneity is substantial: the small-large F1 gap ranges from 1.8 (focal length) to 14.3 (camera) F1 points, so uniform routing leaves cost savings unused.

5. Related Work

Calibration and proper scoring rules (Gneiting & Raftery, 2007; DeGroot & Fienberg, 1983) underpin probabilistic forecasting. ECE and reliability diagrams are standard diagnostics in neural calibration (Guo et al., 2017). Cost-aware LLM inference includes FrugalGPT (Chen et al., 2024a), confident adaptive language modeling (Schuster et al., 2022), and speculative cascades (Chen et al., 2024b). Selective classification (Geifman & El-Yaniv, 2017) and learning to defer (Madras et al., 2018; Mozannar & Sontag, 2020) share the constrained-decision framing. UCCI differs by making the routing score a calibrated forecast and evaluating the operational decision induced by that forecast.

6. Discussion

This paper argues that forecasting model success is a useful primitive for AI systems. The forecast target is not a weather event or market outcome; it is whether an AI component will succeed moments from now. But the same principles apply: probabilities should be calibrated, forecasts should be scored, and decisions should be optimized against the cost of acting on the forecast. For AI forecasting, that makes cascaded inference a tractable setting where probabilistic reasoning, scoring rules, and decision theory directly change system cost.

Limitations. The evidence is from one NER domain; calibration is static; costs are measured as latency rather than dollars or energy. Online recalibration and shift-aware forecasting are natural next steps.

Reproducibility and disclosure. Hyperparameters, splits, prompt template, and serving setup are listed here and in Section B. Raw production queries cannot be released for privacy reasons; calibration, threshold-selection, and evaluation scripts will be released. The supplementary material contains an anonymized longer version of this work for reviewer context.

References

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

Barlow, R. E., Bartholomew, D. J., Bremner, J. M., and Brunk, H. D. *Statistical Inference Under Order Restrictions: The Theory and Application of Isotonic Regression*. John Wiley & Sons, 1972.

Chen, L., Zaharia, M., and Zou, J. Frugalgpt: How to use large language models while reducing cost and improving performance. *Transactions on Machine Learning Research*, 2024a.

Chen, Z., Yang, X., Lin, J., Sun, C., Chang, K. C.-C., and Huang, J. Cascade speculative drafting for even faster LLM inference. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024b.

DeGroot, M. H. and Fienberg, S. E. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 32(1/2):12–22, 1983.

Geifman, Y. and El-Yaniv, R. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Gneiting, T. and Raftery, A. E. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017.

Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C. H., Gonzalez, J. E., Zhang, H., and Stoica, I. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles (SOSP)*, 2023.

Madras, D., Pitassi, T., and Zemel, R. Predict responsibly: Improving fairness and accuracy by learning to defer. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Mozannar, H. and Sontag, D. Consistent estimators for learning to defer to an expert. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020.

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J. V., Lakshminarayanan, B., and Snoek, J. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Sadinle, M., Lei, J., and Wasserman, L. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525):223–234, 2019.

Schuster, T., Fisch, A., Gupta, J., Dehghani, M., Bahri, D., Tran, V. Q., Tay, Y., and Metzler, D. Confident adaptive language modeling. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Zadrozny, B. and Elkan, C. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2002.

A. Proofs

A.1. Proof of Theorem 3.1

Let π be any policy that depends on $u(x)$ and satisfies the accuracy constraint. Escalating query x replaces the small model's expected accuracy $1 - \hat{p}(x)$ with a_ℓ at additional cost $c_\ell - c_s > 0$. The marginal accuracy benefit is $\Delta(x) = a_\ell - 1 + \hat{p}(x)$, which is monotone non-decreasing in $\hat{p}(x)$. To minimize cost while reaching target τ, the policy should buy marginal accuracy in decreasing order of $\Delta(x)/(c_\ell - c_s)$, equivalently decreasing order of $\hat{p}(x)$, until the constraint binds. The smallest threshold satisfying the empirical or population constraint implements this ordering. Since policies in the theorem are restricted to depend only on $u(x)$ and $\hat{p} = g(u)$, this is the optimal policy in the class. □

A.2. Proof of Theorem 3.2

Isotonic regression is a non-parametric maximum-likelihood estimator for monotone functions. Under bounded-density and monotonicity conditions, $\mathbb{E}\|\hat{g} - g^\star\|_2^2 = O(n^{-2/3})$ (Barlow et al., 1972). ECE is an L_1 functional of calibration

error, so Cauchy-Schwarz gives ECE $\leq \sqrt{\mathbb{E}\|\hat{g} - g^\star\|_2^2} = O(n^{-1/3})$. Concentration of empirical ECE around its expectation follows from standard empirical-process bounds. □

B. Reproducibility details

Prompt.

```
Extract entities from this photo
search query. Return JSON with fields:
camera, lens, aperture, shutter_speed,
iso, focal_length.

Query: [USER_INPUT]
Output:
```

Hyperparameters. Temperature 0; max tokens 256; vLLM 0.4.2; H100 80GB; greedy decoding. Calibration: scikit-learn isotonic regression, default settings; trained on 22,500 examples. Threshold selection: grid over $\theta \in [0, 1]$ at resolution 0.005 on the validation set; choose the lowest-cost threshold with validation F1 at least τ.

Cost measurement. Mean end-to-end latency over 100 cold-cache queries: small 47.2 ± 3.1 ms, large 142.3 ± 8.7 ms; ratio 3.02.

Bootstrap CIs. 1000 resamples over test queries.