Cascaded Language Models for Cost-effective Human-AI Decision-Making

UNIVERSITY OF CAMBRIDGE

van_der_Schaar \ LAB

Claudio Fanconi Mihaela van der Schaar

Motivation

Challenge: Deploying large language models in high-stakes settings (e.g. healthcare, finance, education) requires balancing :

- **accuracy**, **cost**, and **reliability**.

Problem:

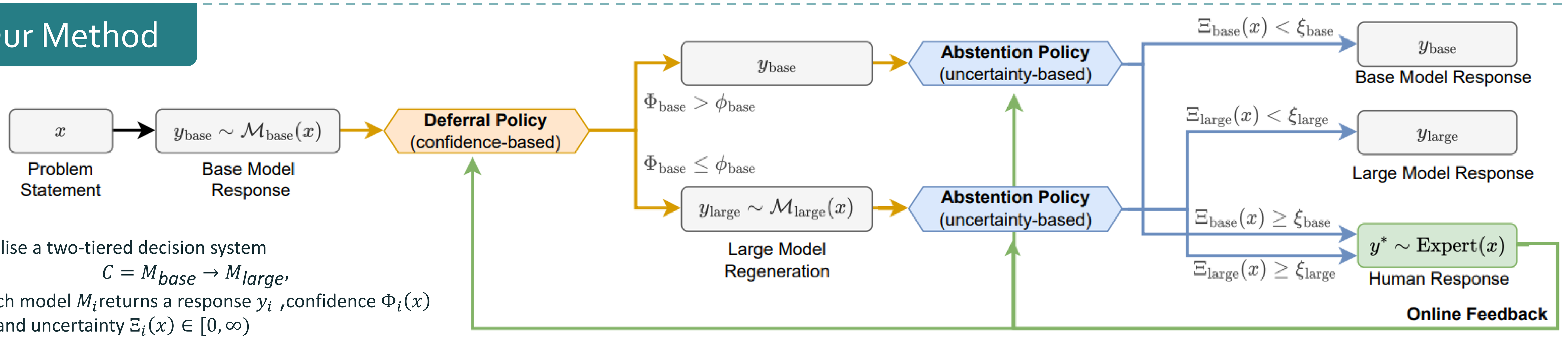
- Larger models offer superior reasoning but incur high computational cost.
- Smaller models are cheaper but struggle on complex tasks.
- Current systems lack mechanisms to decide when to defer to stronger models or humans.

Goal: Build AI systems that make cost-aware and risk-sensitive decisions, learning *when to trust themselves, when to escalate to stronger models, and when to abstain for human review.*

Key Idea:
A cascaded human-AI decision framework that:

- Uses a base model for low-complexity queries.
- Defers to a more capable model when uncertainty is high.
- Abstains and seeks human input when confidence remains low.
- Learns online from feedback to improve over time.

Our Method



We formalise a two-tiered decision system

$$C = M_{base} \to M_{large},$$

where each model M_i returns a response y_i ,confidence $\Phi_i(x) \in [0,1]$, and uncertainty $\Xi_i(x) \in [0, \infty)$

The cascade selects the output according to:

$$C(x) = \begin{cases} M_{base}(x), & \Phi_{base}(x) > \phi_{base},\ \Xi_{base}(x) < \xi_{base} \\ M_{large}(x), & \Phi_{base}(x) \le \phi_{base},\ \Xi_{large}(x) < \xi_{large} \\ \emptyset, & \textit{otherwise (abstain to human)} \end{cases}$$

The **system risk** jointly optimises for accuracy, computational cost, and abstention:

$$R(C) = P(error \wedge \neg abstention) + \lambda_c\, \mathbb{E}[Cost] + \lambda_a\, P(abstention)$$

1 Efficient Deferral

Only defer when $\Phi_{base}(x)$is low to minimise unnecessary regeneration cost:
M_{large} *used iff* $\Phi_{base}(x) \le \phi_{base}$

2 Safe Abstention

Escalate to human experts only when model uncertainty is high.
Abstain *iff* $\Xi_i(x) > \xi_i$

3 Online Adaption

Update thresholds $\phi_{base}, \xi_{base}, \xi_{large}$via stochastic gradient descent on $R(C)$ incorporating feedback from human-labelled abstentions.

$$\theta_{t+1} = \theta_t - \eta \nabla_\theta R(C)$$

Verification & Uncertainty Estimation

Verification Strategies
Given an input–response pair (x, y_i) from model M_i:
- Self-Verification (SV):
M_i re-evaluates its own answer with a verification prompt.
- Surrogate Token Probability (STP):
Extract next-token probability for YES/NO with verification question:

$$p_i(x) = \frac{M_i(\text{YES} \mid x, y_i)}{M_i(\text{YES} \mid x, y_i) + M_i(\text{NO} \mid x, y_i)}.$$

Bayesian Calibration
We fit a **Bayesian Logistic Regression** on a small calibration set (=100 samples):

$$\Phi_i(x) = \mathbb{E}[correctness],$$
$$\Xi_i(x) = STD[correctness],$$

yielding calibrated
- **confidence** $\Phi_i(x)$
- **uncertainty** $\Xi_i(x)$

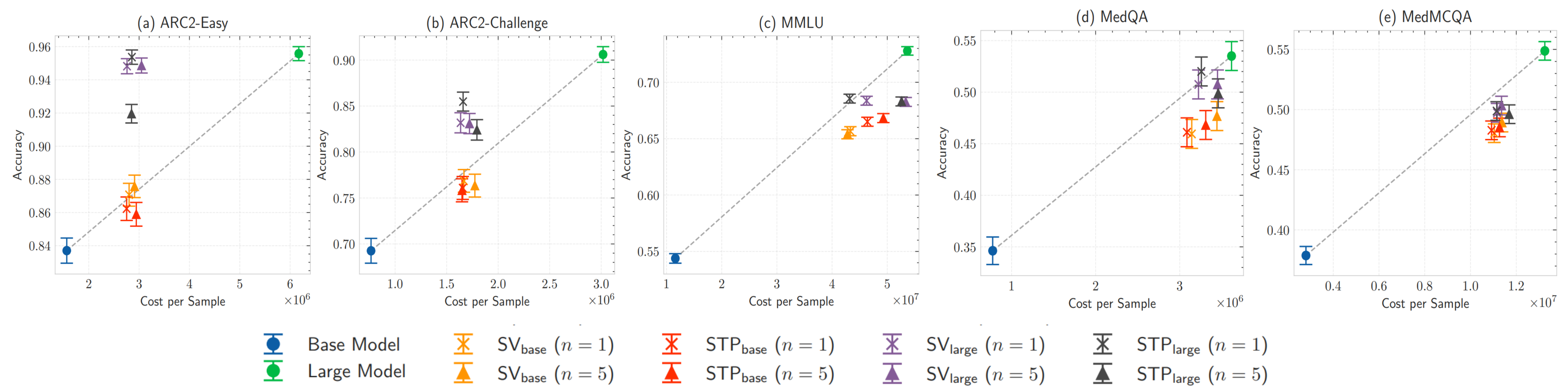
Deferral: if $\Phi_{base}(x) < \phi_{base}$
Abstention: if $\Xi_i(x) > \xi_i$

💡 *Purpose:* This module converts model outputs into **calibrated probabilities** that the system uses to decide **when to defer or abstain**, making the cascade statistically grounded and cost-aware.

Results & Insights

Verification Experiment: Large-model evaluation improves cost–accuracy trade-offs

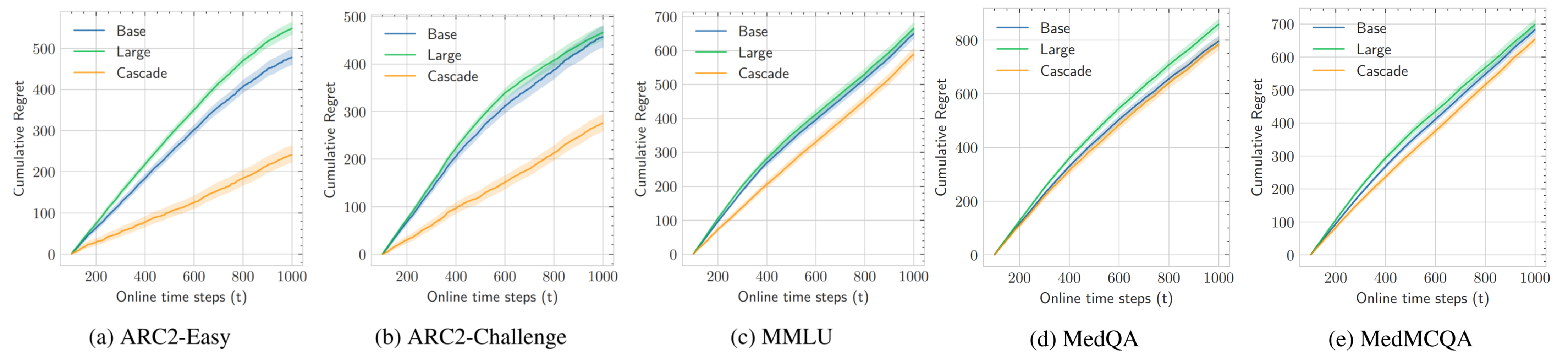
- When using large-on-small model verification the cascaded system achieves higher accuracy per unit cost.
- Surrogate Token Probability (STP) yields the best balance across datasets.



💡 Verifying with a stronger model is cheaper than regenerating responses and yields better calibration.

Online learning Experiment: Online learning reduces cumulative regret

- Introduce a soft-gate (sigmoid) for the decision thresholds, so the optimisation becomes differentiable.
- Updating thresholds $\phi_{base}, \xi_{base}, \xi_{large}$via SGD on $R(C)$ improves decision efficiency over time.



💡 The cascade learns when to defer or abstain, achieving lower long-term cost–risk.

Read the full paper:

Enjoyed this work? Let's connect:

caf83@cam.ac.uk
in/ claudio-fanconi
fanconic.github.com

Acknowledgement: Canon Medical Systems Corporation funds CF's studentship

Canon
CANON MEDICAL SYSTEMS CORPORATION