Gatekeeper: Improving Model Cascades Through Confidence Tuning

Stephan Rabanser[1] Nathalie Rauschmayr[2] Achin Kulshrestha[2]
Petra Poklukar[2] Wittawat Jitkrittum[2] Sean Augenstein[2]
Congchao Wang[2] Federico Tombari[2]

[1]University of Toronto & Vector Institute [2]Google

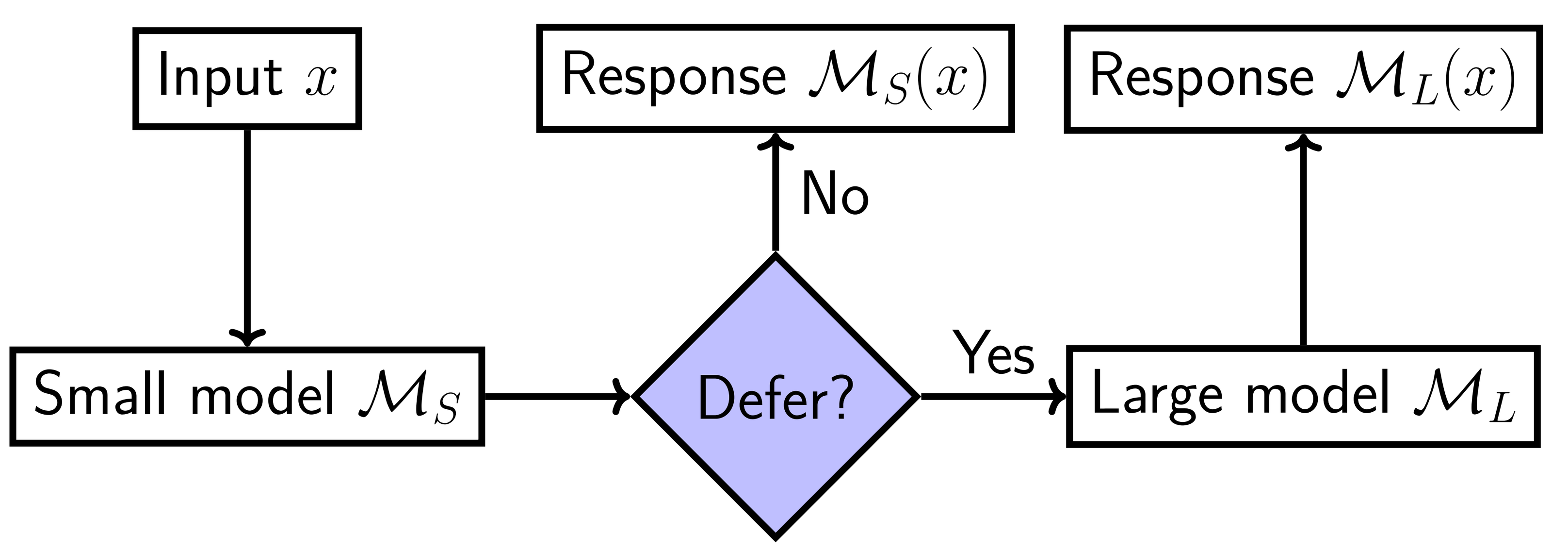
University of Toronto · Vector Institute · Google · ICML International Conference On Machine Learning

Main Contribution

We introduce a new loss function **that calibrates smaller models in cascade setups** to confidently handle easy examples while deferring complex ones.

Paper:

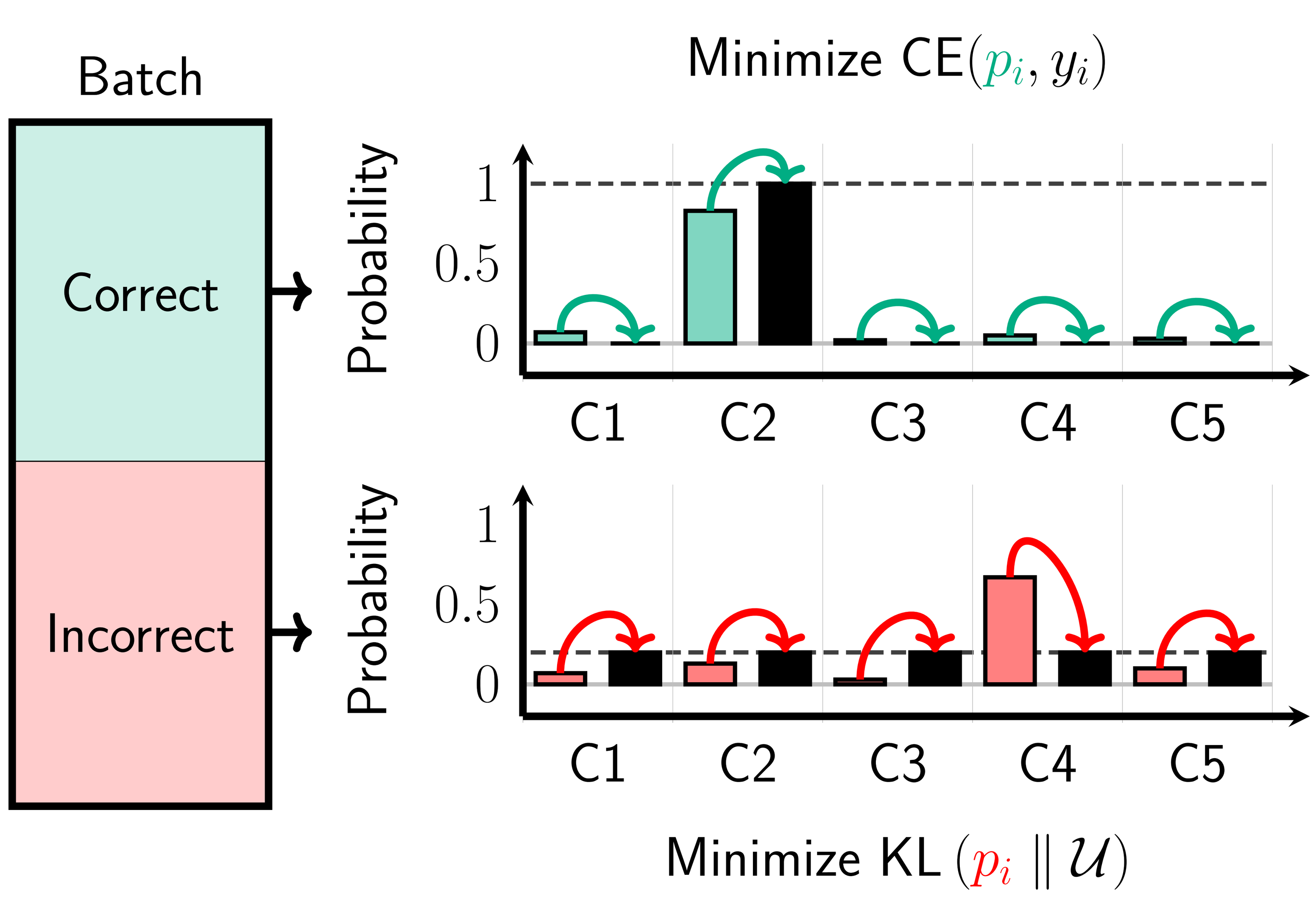
Model Cascading Overview



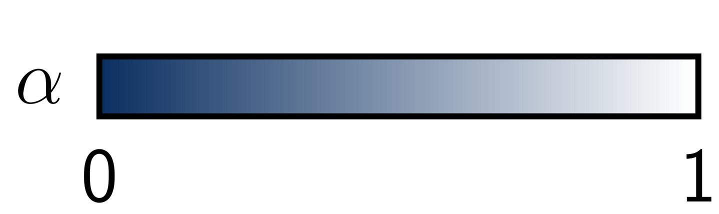
The Gatekeeper Loss

Idea

Fine-tune the small model $\mathcal{M}_S$ by regularizing wrong predictions to a uniform distribution.



$$\mathcal{L} = \alpha\mathcal{L}_{\text{corr}} + (1-\alpha)\mathcal{L}_{\text{incorr}}$$



When optimizing this loss, low α emphasizes confidence calibration of incorrect data points; high α emphasizes maintaining high utility over full distribution.

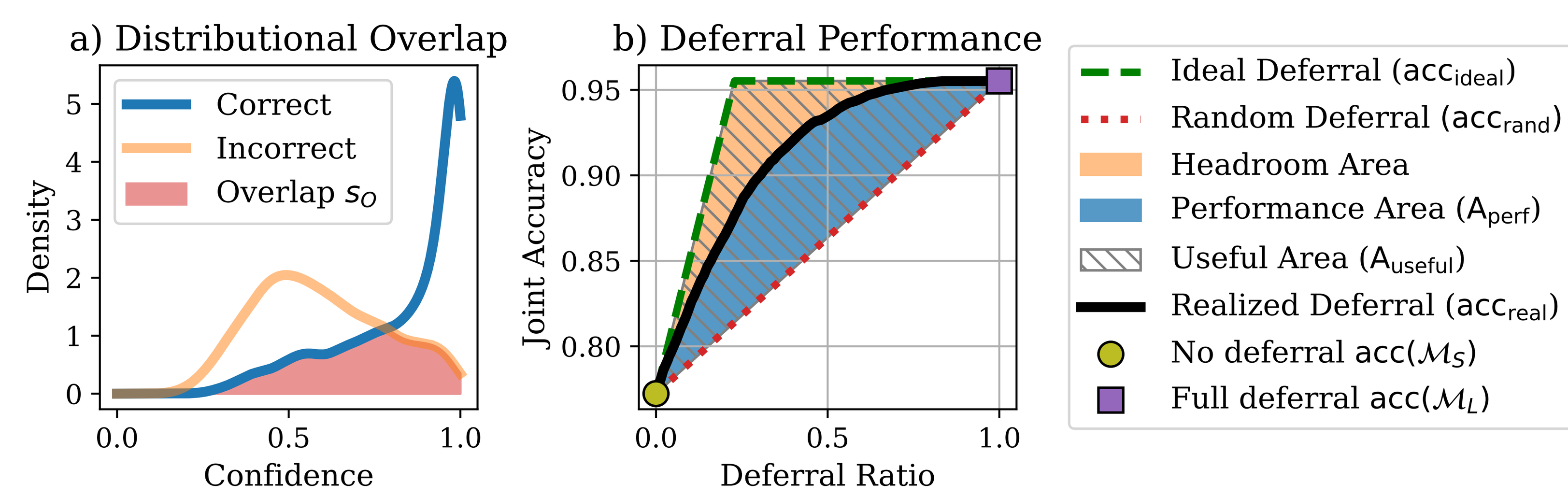
The loss terms take the following form:

$$\mathcal{L}_{\text{corr}} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\{y_i = \hat{y}_i\}\,\text{CE}(p_i, y_i)$$

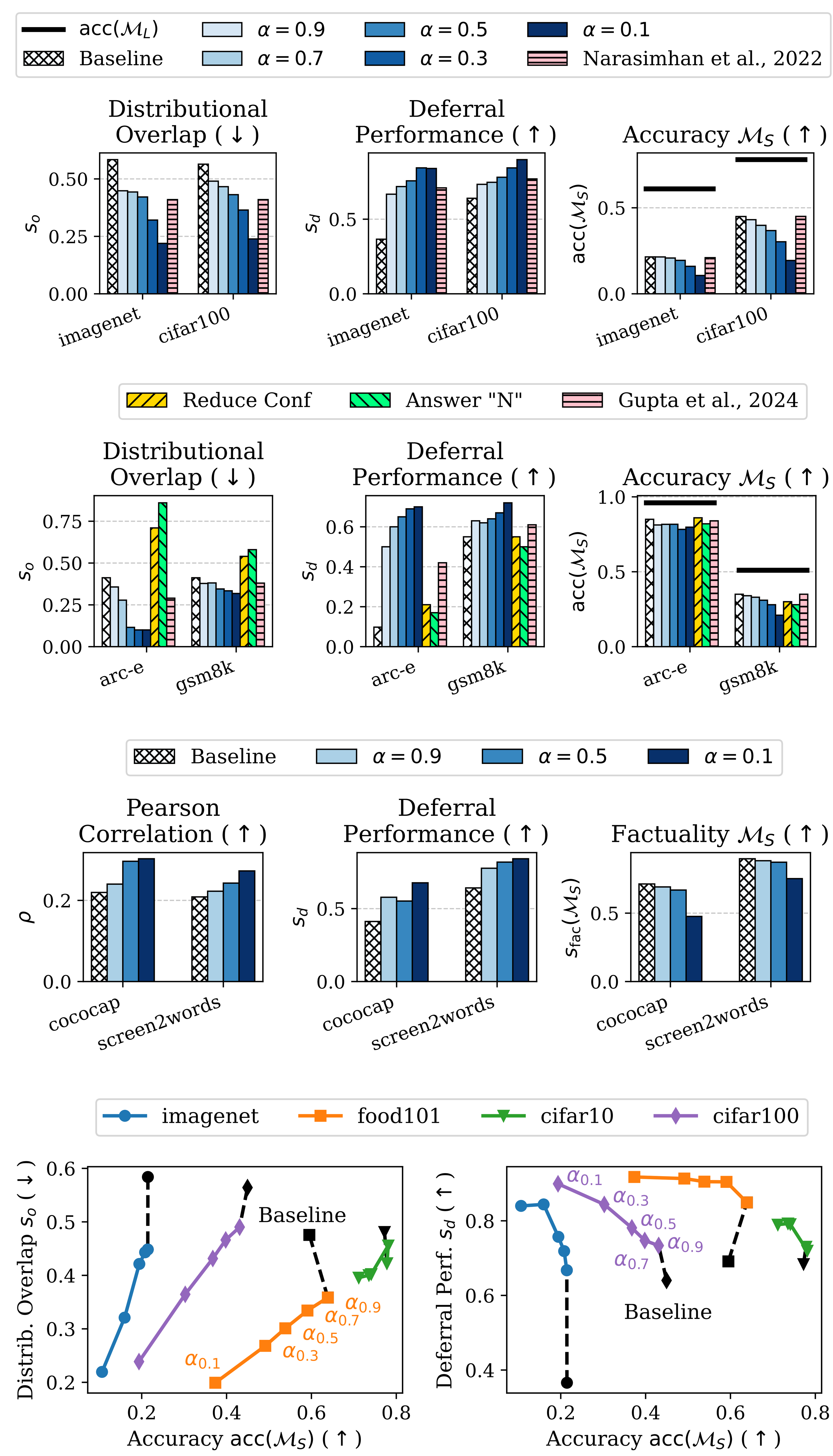
$$\mathcal{L}_{\text{incorr}} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\{y_i \neq \hat{y}_i\}\,\text{KL}(p_i \parallel \mathcal{U})$$

Analogous extension to token-based models possible.

Evaluation Metrics



Experimental Results



Insight

Across all modalities, Gatekeeper enables improved deferral performance by better separating correct versus incorrect predictions, especially at low α. However, this comes at the cost of reduced small model accuracy.