Classifier Uncertainty Beyond Calibration

Sébastien Melo, Gaël Varoquaux, and Marine Le Morvan.

Why Reliable Confidence Scores Matter

The Need for Trustworthy Confidence

- For black-box models in high-stakes settings: need to know the uncertainty behind a decision = **confidence scores**.

Why Reliable Confidence Scores Matter

The Need for Trustworthy Confidence

- For black-box models in high-stakes settings: need to know the uncertainty behind a decision = **confidence scores**.
- Simple metrics (e.g. accuracy) aren't enough: need to evaluate the probabilistic quality of the confidence scores.

Why Reliable Confidence Scores Matter

The Need for Trustworthy Confidence

- For black-box models in high-stakes settings: need to know the uncertainty behind a decision = **confidence scores**.
- Simple metrics (e.g. accuracy) aren't enough: need to evaluate the probabilistic quality of the confidence scores.

Decomposing Uncertainty We measure $d(f(X), Y)$:

- Contains the randomness of the task

Why Reliable Confidence Scores Matter

The Need for Trustworthy Confidence

- For black-box models in high-stakes settings: need to know the uncertainty behind a decision = **confidence scores**.
- Simple metrics (e.g. accuracy) aren't enough: need to evaluate the probabilistic quality of the confidence scores.

Decomposing Uncertainty We measure $d(f(X), Y)$:

- Contains the randomness of the task
- The distance between $f(X)$ and $\mathbb{P}[Y = 1 \mid X]$ too!

Our Results: Better Estimators for Better Decisions

1. More Sample-Efficient Estimators

- We introduce binning-free estimators for the grouping loss and its associated decision risk.

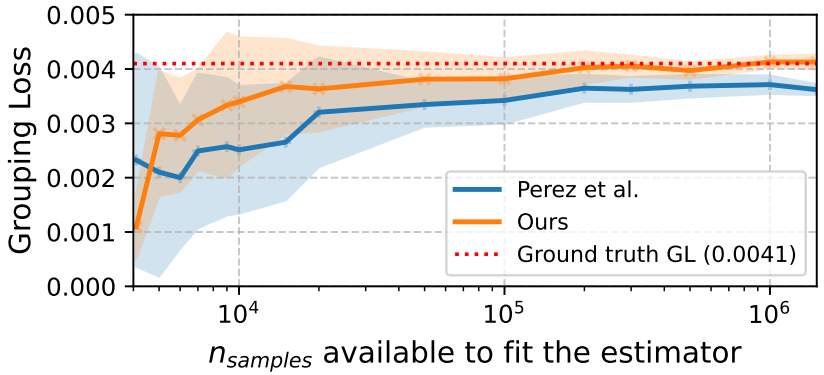

Figure 1: Our estimator converges faster and more tightly than prior work.

Our Results: Better Estimators for Better Decisions

1. More Sample-Efficient Estimators

- We introduce binning-free estimators for the grouping loss and its associated decision risk.

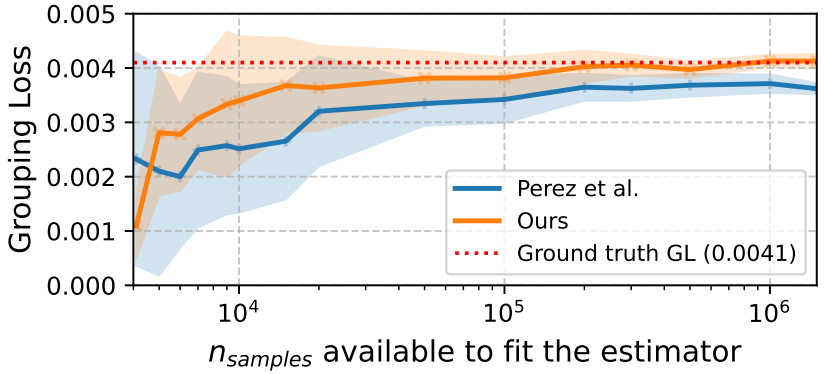

Figure 1: Our estimator converges faster and more tightly than prior work.

2. Improved Individual Decisions with LLM Cascades

- We use our risk estimates as a per-query quality score to build intelligent LLM cascades.

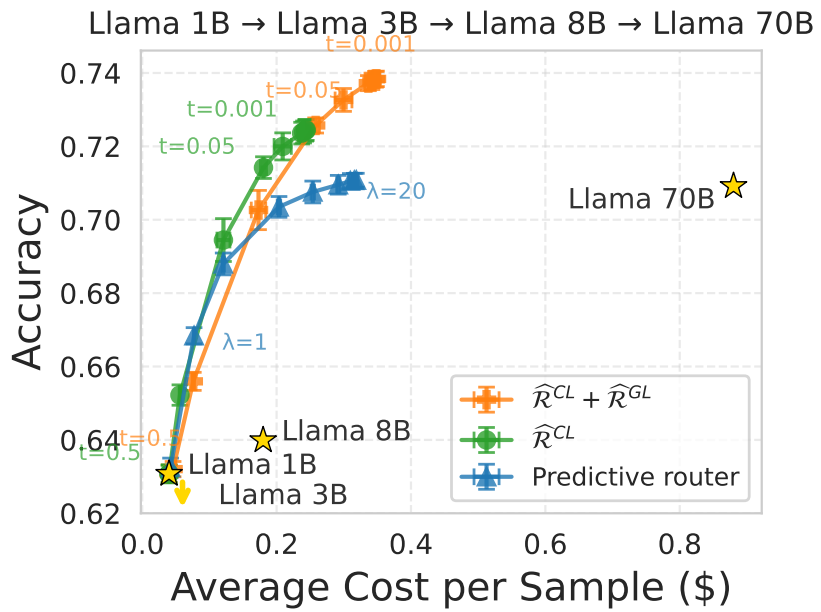

Figure 2: Our cascade improves accuracy while reducing cost compared to baselines.