WHITE PAPER / MODEL EVALUATION

When is a score good enough to act on?

A classifier's first choice can stay exactly the same while its confidence scores send a very different set of cases to reviewers. We measured both effects in a document-reading experiment.

The decision behind the score

Imagine a form-processing queue that sends a digit to an operator whenever the model's confidence falls below 95%. Improving the model's accuracy is one way to improve the queue. A second question is whether a score near 95% has earned the meaning that the routing rule gives it.

The ICLR study *A Benchmark Study on Calibration* examines how choices of model and metric affect calibration. We used that question to design a small, repeatable test. We held the classifier fixed, adjusted the scale of its scores, then measured the effect on a queue with a fixed acceptance threshold.

One model, three independent splits

We used 1,797 labeled, eight-by-eight pixel images from the public handwritten digits dataset. With a fixed random seed of 42 and stratification by digit, we assigned 1,078 images to training, 359 to calibration, and 360 to a final test. Each intensity was divided by 16 before fitting a logistic regression classifier.

We divided the classifier's scores by a single positive number, the temperature, before converting them to probabilities. We fitted that number on the calibration split by minimizing log loss. The fitted temperature of 0.567 sharpened the probabilities and preserved the ordering of the ten possible digits.

A smaller queue carries a different cost

Held-out results · 360 digits · 95% acceptance threshold

Measure	Original	Calibrated
Correct first choice	344 / 360	344 / 360
Log loss (lower is better)	0.189	0.131
Cases sent for review	164	70
Errors automatically accepted	0 of 196	1 of 290

At the same threshold the revised scores would send 94 fewer digits for review and allow one additional incorrect digit through. Log loss improved even though every first-choice prediction stayed the same. A team looking only at top-choice accuracy would miss both changes.

How to choose a threshold

Estimate the time an operator spends checking a digit and the downstream cost of accepting an incorrect one. In this test, the decision turns on whether avoiding 94 reviews is worth one additional error. A field used to route a package can have a very different cost from a field used to reconcile a payment.

Before adopting a threshold, repeat the measurement on recent documents, split by capture device, field type, and input quality. Handwritten benchmark digits are small, clean crops. Their measured queue sizes give a reproducible illustration, while the incoming document mix determines the operating point.

Research & data

- A Benchmark Study on Calibration · ICLR 2024
- Optical recognition of handwritten digits · scikit-learn dataset