

Evaluating and Recalibrating Probabilistic Classifiers from One-Bit Threshold Queries

How much probabilistic evaluation is possible when probabilities are hidden?

Roman Plaud Antoine Saillenfest Matthieu Labeau
Thomas Bonald Willem Waegeman

onepoint.



Can we evaluate probabilities we are not allowed to see?

A binary example

Question: Is the Pacific Ocean larger than the Atlantic Ocean?

Ground truth: Yes



The model behaves like a binary classifier, but the API may reveal only the final token.

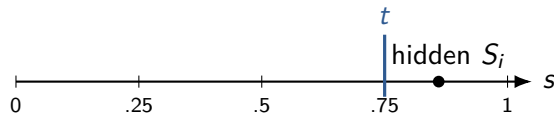
hidden score $S_i \in (0, 1)$

e.g. $S_i = 0.73$
but we do not observe it

Standard calibration and re-calibration methods normally assume that S_i is available.

Can we still do probabilistic evaluation without reconstructing the hidden scores?

A black box can still answer a comparison question



$$B_i = \mathbf{1}\{S_i \geq T_i\}$$

For each labeled example, we choose one threshold T_i and observe one bit B_i .

In an LLM API

Restrict the answer to two tokens, say A/B. A suitable `logit_bias` shifts the decoding boundary so that the returned token tells us whether $S_i \geq t$.

choose threshold t
apply corresponding logit bias

one API call

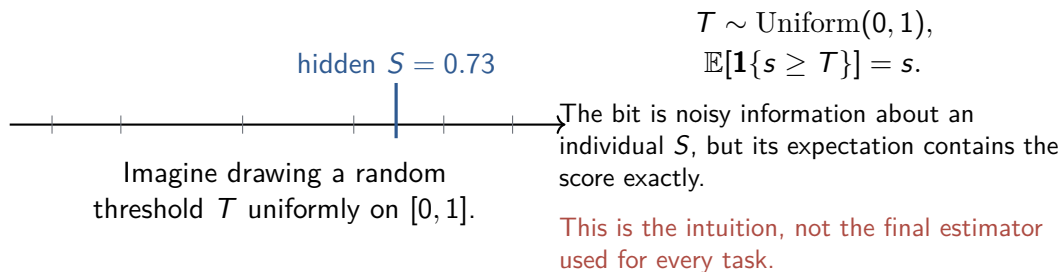
observe A or B
 $\iff$ one threshold bit

Constraint: at most one threshold query per example.

The key idea: do not reconstruct each probability

One bit about one score tells us very little.

One bit about many scores, at chosen/random thresholds, can reveal population quantities.



Paper's strategy: express the quantity of interest in terms of threshold events, then estimate it directly from those events.

What can one bit per example buy us?

1. Calibration error

How far are predicted probabilities from empirical frequencies?

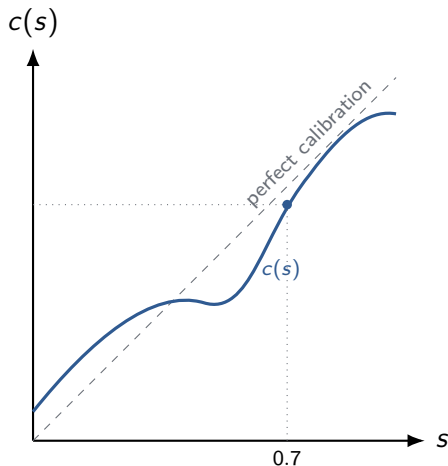
2. Proper scoring rules

Can we estimate Brier score / Log-loss without seeing probabilities?

3. Recalibration

Can we fit Temperature Scaling from the same one-bit data?

1. Calibration: what are we trying to measure?



If a model predicts 70% on many examples, we would like roughly 70% of them to be positive.

Define the **calibration function**

$$c(s) = \mathbb{E}[Y \mid S = s].$$

Perfect calibration means

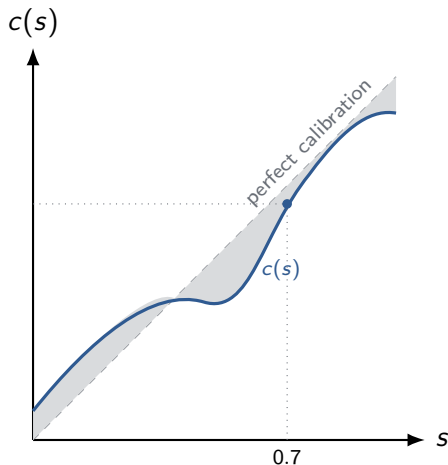
$$c(s) = s.$$

The paper measures the average deviation from this ideal:

$$\text{TCE} = \mathbb{E}[|c(S) - S|].$$

Goal: estimate TCE without ever observing S .

1. Calibration: what are we trying to measure?



If a model predicts 70% on many examples, we would like roughly 70% of them to be positive.

Define the **calibration function**

$$c(s) = \mathbb{E}[Y \mid S = s].$$

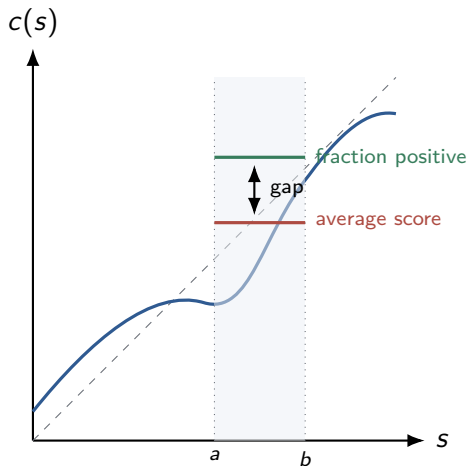
Perfect calibration means

$$c(s) = s.$$

The paper measures the average deviation from this ideal:

$$\text{TCE} = \mathbb{E}[|c(S) - S|].$$

1. Calibration: with full probabilities, estimate it locally



We cannot estimate $c(s)$ separately at every exact score.

So we group nearby scores into bins.

Inside one bin

calibration gap

$\approx$

fraction positive - average score

Aggregate these local gaps across the score distribution to estimate TCE.

But under one-bit access, we know neither the score S_i nor which bin contains it.

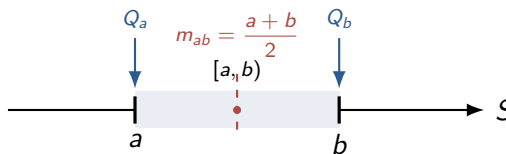
1. Calibration: from hidden bins to threshold queries

The hidden score S is not assigned to a visible bin

$$\mathbf{1}\{S \in [a, b)\} = \mathbf{1}\{S \geq a\} - \mathbf{1}\{S \geq b\}$$

Inside the bin, the score value is still hidden

$$S \approx m_{ab} := \frac{a + b}{2} \quad \text{for } S \in [a, b)$$

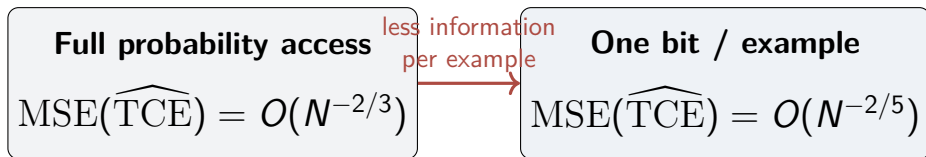


One-query constraint

Use one subset for boundary a and another subset for boundary b .

Then estimate the bin contribution by the **difference of empirical averages**.

1. Calibration: how accurately can we estimate TCE?



Optimal bin resolutions: $M \asymp N^{1/3}$ with full access, and $M \asymp N^{1/5}$ with one-bit access.

Same target. Still consistent. More data needed.

2. Proper scoring rules: what do we want to estimate?

We want to measure the model's **population predictive risk**:

$$R_\phi(f) = \mathbb{E}[\phi(S, Y)]$$

where ϕ can be, for example, the Brier score or Log-loss.

With full probability access, we observe S_i and estimate

$$\hat{R}_{\text{full}} = \frac{1}{N} \sum_{i=1}^N \phi(S_i, Y_i).$$

But under one-bit access, S_i is hidden.

Can we estimate the same population quantity using only

$$B_i = \mathbf{1}\{S_i \geq T_i\} ?$$

2. Proper scoring rules from one threshold query

Binary proper scores can be written as

$$\phi(s, y) = \int_0^1 w(t) L_t(s, y) dt$$

where L_t only needs

$$\mathbf{1}\{s \geq t\}.$$

Sample

$$T \sim \text{Unif}(0, 1), \quad B = \mathbf{1}\{S \geq T\}.$$

Then

$$\hat{\phi}_{1\text{bit}} = w(T) L_T(S, Y)$$

is computable from (T, B, Y) , and

$$\mathbb{E}_T[\hat{\phi}_{1\text{bit}} \mid S, Y] = \phi(S, Y).$$

Thus

$$\hat{R}_{1\text{bit}} = \frac{1}{N} \sum_{i=1}^N \hat{\phi}_{1\text{bit}, i}$$

is unbiased for $R_\phi(f)$.

Same population risk, without observing the probabilities.

3. Recalibration: monotonicity lets us reuse the same bits

Temperature Scaling applies a monotone transformation q_α to the hidden score.

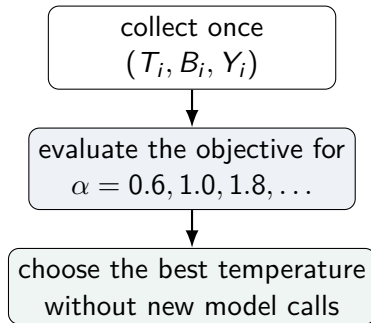
Because q_α is strictly increasing,

$$\mathbf{1}\{q_\alpha(S) \geq q_\alpha(t)\} = \mathbf{1}\{S \geq t\}$$

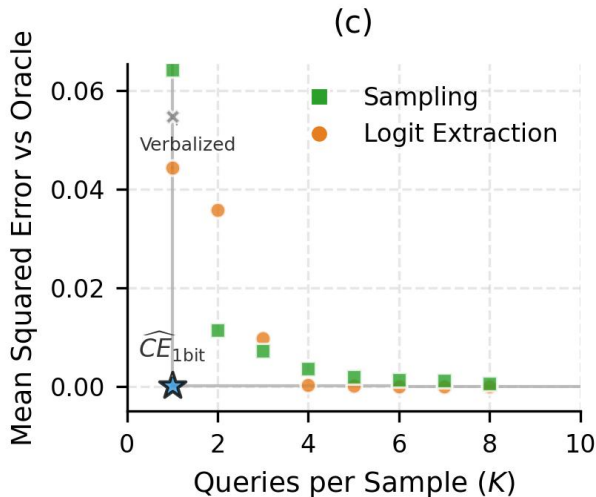
So the bit we already observed at threshold t contains the threshold information needed for **every candidate temperature** α .

Practical consequence

Once (T_i, B_i, Y_i) has been collected, we can evaluate the one-bit TS objective for many α values **without querying the model again**.



Does one bit suffice in practice? Calibration auditing

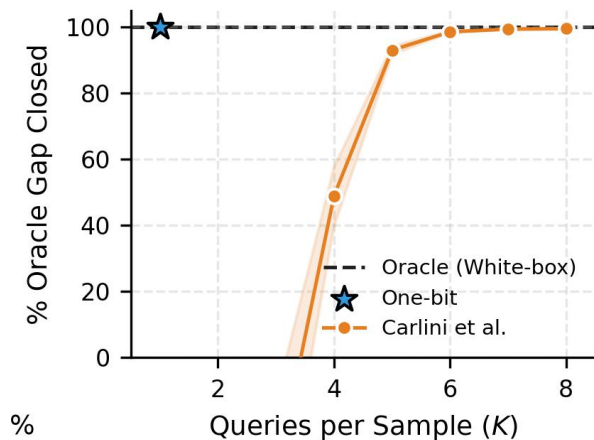


On BoolQ and MMLU across four open-weight instruction-tuned LLMs:

- ▶ one-bit estimation uses **one query per example**;
- ▶ it substantially outperforms verbalized confidence at that budget;
- ▶ logit extraction needs roughly **4–5 queries/example** to reach comparable calibration-estimation error.

Does one bit suffice in practice? Temperature Scaling

(d) LLM Recalibration (Avg)



$> 98\%$

of the probability-access
Temperature Scaling
improvement is recovered on
average with

1 query/example

Carlini-style logit extraction needs
about **5–6 queries/example** for
comparable performance.

Takeaway

The paper is not about extracting hidden probabilities cheaply.

It shows that, for several probabilistic tasks,
we never needed to extract them in the first place.

Calibration error

Consistent from threshold
differences

Price: slower rate

Proper scores

Unbiased via Monte Carlo
threshold sampling

Price: extra variance

Recalibration

Fit TS from the same one-bit
data

Need: centered objective

Open questions: multiclass extensions; optimal one-bit calibration rates; adaptive threshold allocation; interfaces that expose suitable threshold interventions.

Backup: formal one-bit calibration estimator

Let $B_m = [t_{m-1}, t_m)$ and $a_m = (t_{m-1} + t_m)/2$. Split the data into threshold groups $\mathcal{I}_0, \dots, \mathcal{I}_{M-1}$. Then

$$\hat{\Delta}_m^{\text{1bit}} = \frac{1}{|\mathcal{I}_{m-1}|} \sum_{i \in \mathcal{I}_{m-1}} (Y_i - a_m) Q_{t_{m-1}}(X_i) - \frac{1}{|\mathcal{I}_m|} \sum_{i \in \mathcal{I}_m} (Y_i - a_m) Q_{t_m}(X_i).$$

The aggregate estimator is

$$\widehat{\text{CE}}_M^{\text{1bit}} = \sum_{m=1}^M \left| \hat{\Delta}_m^{\text{1bit}} \right|.$$

Under the paper's regularity assumptions, balancing M^3/N and M^{-2} yields $M \asymp N^{1/5}$ and MSE upper bound $O(N^{-2/5})$.

Backup: one-bit proper-score estimator

For a binary proper score with

$$\phi(s, y) = \int_0^1 w(t) L_t(s, y) dt,$$

draw $T \sim g$ and observe $B = \mathbf{1}\{S \geq T\}$. Since

$$L_T(S, Y) = Y(1 - T)(1 - B) + (1 - Y)TB,$$

define

$$\hat{\phi}_{\text{1bit}}(B, Y, T) = \frac{w(T)}{g(T)} L_T(S, Y).$$

Then

$$\mathbb{E}_T[\hat{\phi}_{\text{1bit}} \mid S, Y] = \phi(S, Y).$$

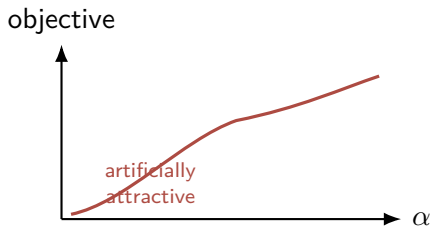
The restriction therefore adds variance, not bias, for fixed proper-score evaluation.

3. Temperature Scaling needs one extra fix: center the objective

Naive one-bit TS objective

For every fixed $\alpha > 0$, the estimator is unbiased.

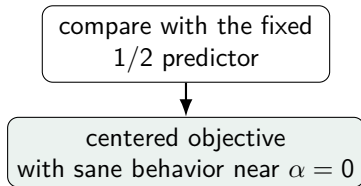
But on a finite dataset its empirical value collapses toward **zero** as $\alpha \rightarrow 0$.



Paper's fix: center at the 1/2 predictor

The true TS risk tends to $\log 2$ as $\alpha \rightarrow 0$, because every score is pushed toward $1/2$.

Subtract the threshold loss of the fixed $1/2$ predictor, estimate the difference from the same bit, then add $\log 2$ back.



Unbiasedness at each fixed α is not, by itself, enough for stable empirical optimization.

Backup: centered one-bit Temperature Scaling objective

For $q_\alpha(s) = \sigma(\alpha \text{logit}(s))$, the centered estimator in the paper is

$$\hat{\ell}_\alpha^{\text{1bit,cent}}(B, Y, T) = \log 2 + \frac{\alpha(q_\alpha(T) - Y)}{g(T)T(1-T)} (B - \mathbf{1}\{T \leq 1/2\}).$$

Once (T, B, Y) is observed, this can be evaluated for any $\alpha > 0$ without further model queries.

Important nuance: the paper constructs an unbiased population objective but does not prove consistency of the empirical minimizer $\hat{\alpha}$; its behavior is assessed empirically.

Backup: experimental setup

- ▶ Benchmarks: BoolQ and MMLU (MMLU converted to binary candidate questions).
- ▶ Models: Qwen-2.5-7B-Instruct, Llama-3.1-8B-Instruct, Mistral-7B-Instruct-v0.3, Gemma-2-9B-IT.
- ▶ Restricted-access baselines: verbalized confidence, sampling frequency, and repeated-query logit extraction.
- ▶ Calibration auditing: one-bit estimator compared against probability-access ECE reference.
- ▶ Temperature Scaling: one-bit TS fitted on calibration data and evaluated by fraction of probability-access TS improvement recovered.

Backup: assumptions and limitations

- ▶ Binary classification setting.
- ▶ At most one threshold query per labeled example.
- ▶ LLM realization assumes binary-answer restriction and greedy decoding.
- ▶ Threshold queries require an interface intervention such as `logit_bias`; not every API exposes one.
- ▶ The $O(N^{-2/5})$ calibration rate is an upper bound for this estimator; whether a faster one-bit estimator exists is open.
- ▶ Multiclass calibration and more adaptive allocation of the query budget remain open directions.