NEURAL INFORMATION PROCESSING SYSTEMS

Proper Scoring Rules for Agentic Uncertainty Quantification

Suresh Raghu[★1] Satwik Pandey[★1] Shashwat Pandey[1]
[1] Independent Researchers · [★] equal contribution

Paper + appendix

MAIN CLAIM **Agentic uncertainty should be evaluated as a probability trace, not only as a ranking, binwise calibration score, or collapsed trajectory summary.**
TPS is a strictly proper trajectory score for $q_t = P(Y = 1 \mid H_t)$, with a censored extension for administratively stopped prefixes.

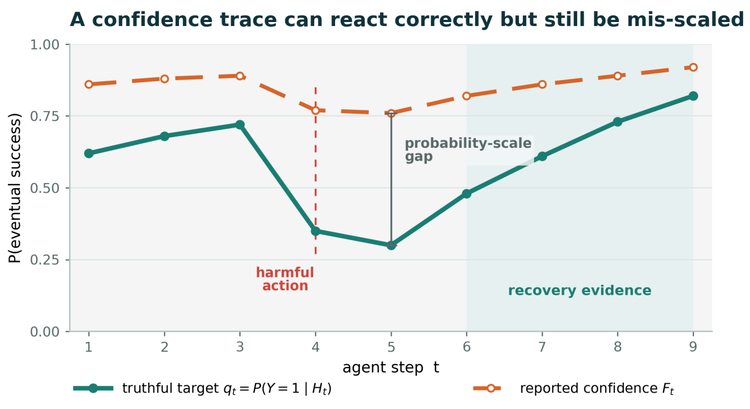
1 What should agentic confidence mean?

Running target. The forecast object is not final confidence; at each prefix H_t, the agent reports F_t as a probability of eventual task success.

history H_t → reported F_t → should mean $q_t = P(Y = 1 \mid H_t)$

The target event Y is fixed: eventual task success. But the information available to the agent changes after every observation, action, and tool result, so q_t can rise or fall over time.

Forecast object: every prefix H_t gets a probability forecast F_t. **Target event:** one fixed terminal outcome Y. **Truthful value:** $q_t = P(Y = 1 \mid H_t)$. The event is fixed; information changes over time. Deferral and human handoff need calibrated probabilities, not just ordering.



Probability scale matters. The reported trace reacts to events, but remains overconfident; deferral and handoff need calibrated probabilities, not just movement direction.

2 What do common evaluators measure?

Evaluator	What it measures	What it can miss
AUROC / AUPRC / AURC	Trajectory ranking	Probability scale
Trajectory ECE	Binwise calibration of a scalar aggregate	Resolution within bins
Scalarized Brier / log	One collapsed trajectory scalar	The full trace
Complete-only evaluation	Completed trajectories only	Censored / harder prefixes
TPS / censored TPS	Full prefix-conditioned process q_t	**Strictly proper target**

Before choosing an evaluator, separate what each metric elicits: ranking, binwise calibration, one scalar summary, or the full probability trace.

Failure proof A: calibration can lose resolution

T-ECE checks bin means after a trace has already been collapsed. A forecast can be calibrated on average but still useless for handoff decisions.

Truthful trace:
$q(a) = 0.2,\ q(b) = 0.8$
same bin average = 0.5

Flat trace:
$G(a) = G(b) = 0.5$
also $T\text{-ECE} = 0$

So binwise calibration alone can pass a trace that destroys within-bin information.

Failure proof B: scalarization changes the target

A proper score remains proper only for the forecast it receives. If the trace is first collapsed to $C = \Phi(F_{1:T})$, Brier/log elicits C, not every prefix.

$$F_{1:T} \to \Phi(F_{1:T}) \to S(C, Y)$$

optimizes the summary, not the probability stream

This is why last/average/min confidence can look good while intermediate F_t is wrong.

3 TPS scores the trace directly

TPS changes the unit of evaluation from one collapsed trajectory number to the adapted probability process itself. Lift *any* strictly proper binary score S onto the whole trace:

$$\mathrm{TPS}(F_{1:T}, Y) = \sum_t w_t S(F_t, Y)$$

Thm 4.1 · expected TPS is uniquely maximized when $F_t = q_t$ at every step

TPS is an evaluator, not a predictor. It evaluates any confidence stream once expressed as $F_t \in [0, 1]$: verbal confidence, token probability, entropy, post-hoc calibrators, or future agentic-UQ methods.

Why strict propriety matters. A proper score rewards honest probabilities. A strictly proper score makes the truthful probability the unique optimum. TPS lifts this idea to trajectories: the best expected score is achieved only when $F_t = q_t$ at every step. This is stronger than ranking trajectories correctly or being calibrated after collapsing the trace.

Theorem 4.1 · Why TPS is strictly proper

At each prefix H_t, the truthful target is $q_t = \mathbb{E}[Y \mid H_t]$.

Because S is a strictly proper binary score, any reported F_t has nonnegative conditional regret:

$$\mathbb{E}[S(q_t, Y) - S(F_t, Y) \mid H_t] \geq 0,$$

with equality only when $F_t = q_t$

TPS just sums these prefix regrets with fixed positive weights:

$$\mathbb{E}[\mathrm{TPS}(q, Y) - \mathrm{TPS}(F, Y)] = \sum_t w_t \, \mathbb{E}[\mathrm{regret}_S(F_t, q_t)] \geq 0.$$

Because every $w_t > 0$, zero total regret forces $F_t = q_t$ at every observed prefix.

TPS therefore elicits the whole probability trace, not just a ranking or a collapsed trajectory score.

What does TPS let the evaluator choose?

WEIGHT SCHEDULE OVER THE TRACE

Weighting	Means	Use when
Front-loaded	early uncertainty matters more	deferral / early handoff
Uniform	every prefix matters equally	general trace reliability
Back-loaded	later recovery matters more	final recovery / late correction

Weights are not predictions. They encode which prefixes downstream decisions care about.

PREDICTOR STREAMS EVALUATED

verbal confidence · completion token prob. · completion entropy · action-span token prob. · action-span entropy · base rate

TPS is predictor-agnostic: any stream becomes evaluable once expressed as $F_t \in [0, 1]$.

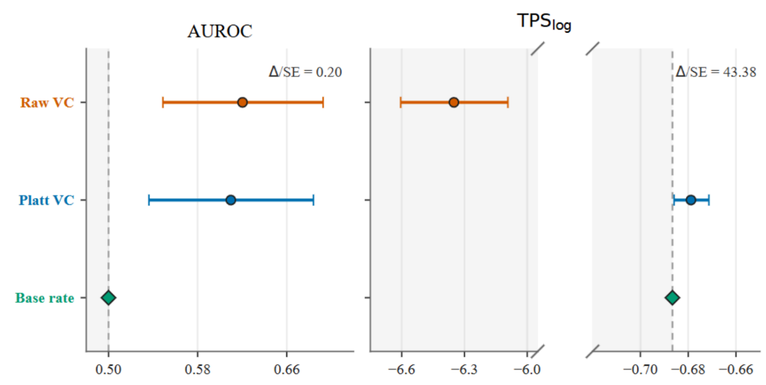
HOW TO READ A TPS RESULT

Reporting choice	What it means	Why it matters
$\mathrm{TPS}_{\log}$	Primary score	Strongly detects confident wrong probabilities
$\mathrm{TPS}_{\mathrm{Brier}}$	Bounded check	Makes sure conclusions are not just log-score harshness
$\mathrm{TPS}_{\beta(2,4)}$	Asymmetric check	Tests cost-shaped sensitivity to different error regimes
Δ/SE	Shift / bootstrap uncertainty	Shows whether evaluator changes are larger than noise
Censored TPS	Scores observed prefixes plus q_Z	Prevents complete-only evaluation from changing the population

Larger TPS is better; for log scores, less negative is better. Score shifts are evaluator effects unless task outcomes themselves change.

4 Trigger 1: same ranking, different probability scale

Tau2-Bench is the clean rank-vs-scale stress test: keep the ordering nearly fixed, change the probability scale, and see which metrics notice.

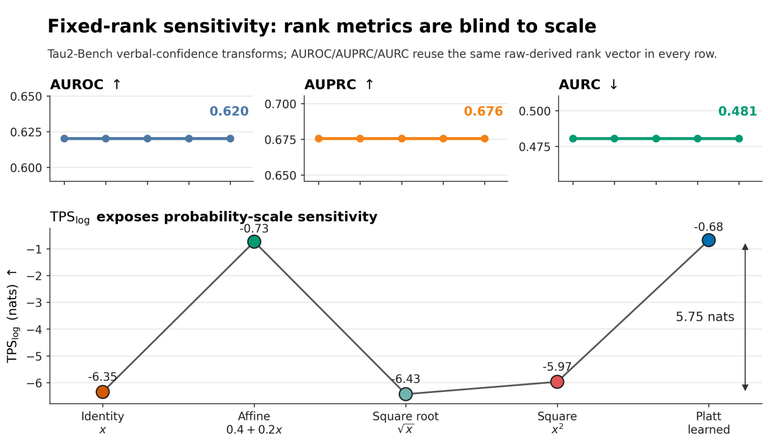


Tau2-Bench. Platt-recalibrating verbal confidence barely moves AUROC, but moves the strictly proper $\mathrm{TPS}_{\log}$ by **tens of bootstrap standard errors**.

All TPS values are reward-oriented: higher is better; for log scores, less negative is better.

0.20
AUROC shift (Δ/SE)
rank metric barely moves

≈ 43
$\mathrm{TPS}_{\log}$ change
Δ/SE 5.67 nats



Fixed rank vector. AUROC/AUPRC/AURC are unchanged by construction, while $\mathrm{TPS}_{\log}$ spans ≈ 5.7 nats because it scores probability scale.

Why censoring matters. Complete-only evaluation changes the population being evaluated. If long or hard trajectories hit a fixed step budget, dropping them can make the reported uncertainty score describe only the easy/completed subset. Censored TPS keeps the observable prefix and states the assumption used for the missing outcome.

5 Mechanism: complete-only drops harder prefixes

WebShop is the population-shift stress test: complete-only evaluation drops exactly the long, hard prefixes where uncertainty matters most.

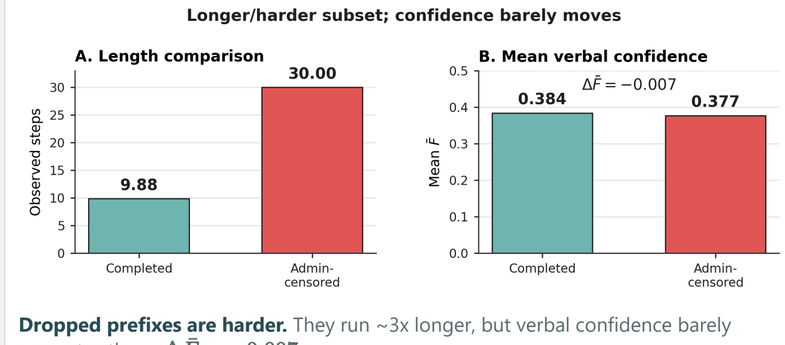
WebShop: 47.08% of valid traces are censored

500 attempted trajectories → 192 parser errors excluded → 308 valid working sample → 163 completed / 145 admin-censored

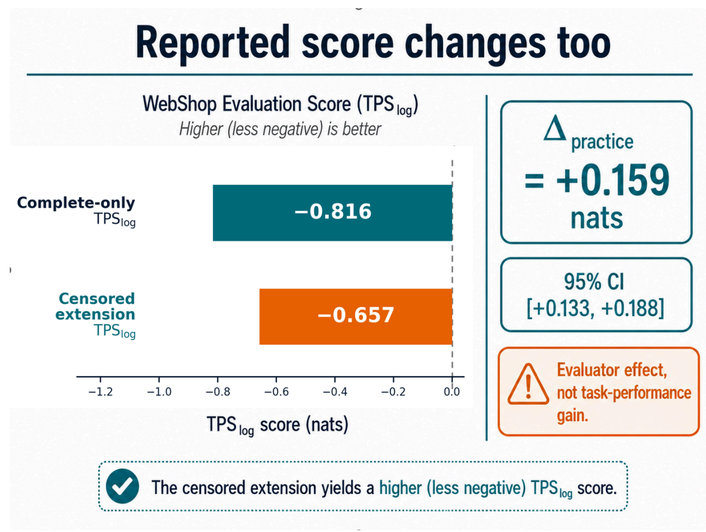
47.08% admin-censored

Complete-only changes the evaluated population. It scores 163 completed trajectories and drops 145 valid max-step prefixes unless censoring is handled.

Complete-only drops the hard subset: 145 max-step prefixes are excluded even though they are valid administrative censored cases.



Dropped prefixes are harder. They run ~3x longer, but verbal confidence barely separates them: $\Delta\bar{F} = -0.007$.



Evaluator effect, not performance gain. Including censored prefixes changes $\mathrm{TPS}_{\log}$ from -0.816 to -0.657: Δ =**+0.159 nats**, 95% CI [+0.133, +0.188].

All TPS values are reward-oriented: higher is better; for log scores, less negative is better.

Sign is stable across predictors: overconfident censored prefixes are penalized, not rewarded.

6 Can q_Z be estimated?

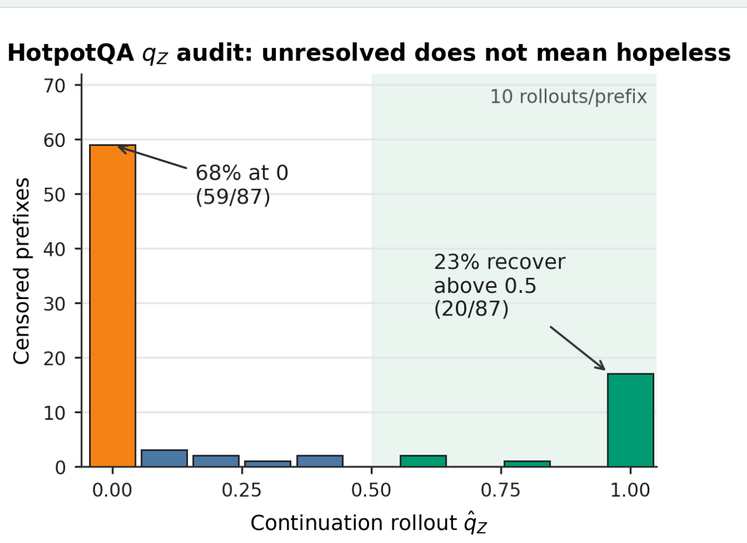
The exact censored score requires a continuation-success probability for the stopped prefix, $q_Z = P(Y = 1 \mid H_Z)$. We audit whether it is operationally estimable.

Censored TPS is a soft-label score

For complete traces, score the observed outcome. For administratively stopped prefixes, score the prefix against a continuation-success weight q_Z.

$$\sum_{t \leq Z} w_t \left[\Delta S(F_t, Y) + (1 - \Delta)\left(q_Z S(F_t, 1) + (1 - q_Z) S(F_t, 0)\right)\right]$$

The assumption is explicit: admin-censoring must not add outcome information beyond the stopped prefix.



Continuation rollouts estimate q_Z. 68% of prefixes remain stuck, but **23% recover above 0.5**; unresolved does not mean hopeless.

How can q_Z be estimated in practice?

For a stopped prefix H_Z, the final outcome Y is unobserved. Censored TPS replaces the missing label with $q_Z = P(Y = 1 \mid H_Z)$.

$$\mathrm{score}(H_Z) = q_Z S(F_t, 1) + (1 - q_Z) S(F_t, 0)$$

So censored TPS is a soft-label score: the stopped prefix is not automatically a failure unless we choose the pessimistic approximation $q_Z = 0$.

Direct continuation rollouts
Resume the stopped prefix M times under the same agent policy.
$\hat{q}_Z^{MC} = M^{-1} \sum_{m=1}^{M} Y^{(m)}$
Best for audits and small datasets.

Plug-in nuisance model
Train $g_\psi(H_Z, Z, \mathrm{metadata}) \approx q_Z$ using completed traces or rollout-audited prefixes.
Best for scalable censored scoring.

Hybrid audit
Roll out a subset of censored prefixes, train g_ψ, then apply it to the full censored set.
Best when full continuation rollouts are expensive.

WHEN IS THIS VALID?
Administrative stops only: once H_Z is known, the stop event should not add extra outcome information. Parser/error stops are separated because they may be informative.

One Object, Three Evaluator Failures, Two Empirical Triggers

Object
Score the prefix-conditioned probability process $q_t = P(Y = 1 \mid H_t)$.

Failure modes
Rank metrics miss scale; T-ECE misses resolution; scalarized losses miss the trace.

Fix
TPS lifts any strictly proper binary score to every prefix with positive weights.

Stress tests
Tau2 exposes scale sensitivity; WebShop exposes complete-only censoring bias.

sureshraghu0706@gmail.com psatwik2711@gmail.com