SYSTEMONEBENCH: A CROSS-DOMAIN BENCHMARK FOR CALIBRATED MACHINE-NATIVE INTELLIGENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Machine intelligence is increasingly embedded in software that requires structured decisions rather than open-ended text generation. Existing evaluations mostly measure generative quality, reasoning accuracy or task completion, and give limited evidence about the calibration, efficiency and reliability of models used as machine-consumed decision functions. System One interfaces, now offered as hosted services, serve this use by returning a probability distribution over a supplied answer space; this guarantees that every answer is well formed, not that it is informative or calibrated. We introduce SystemOneBench, a cross-domain benchmark protocol for calibrated structured decision models, instantiated on seven data families spanning financial market microstructure, blockchain transaction dynamics, road traffic, Earth-system observations, maritime operations and autonomous-agent traces. Each task maps a real-world state and a dynamically defined answer space to typed probabilistic decisions whose ground truth is derived from subsequently observed outcomes rather than model-generated labels. The protocol specifies paired, dependence-aware analyses of decision accuracy, proper scoring rules, calibration error, risk–coverage trade-offs, temporal and entity shift and option-set generalization, and a procedure for measuring latency, throughput and energy. Its harness presents free generation, constrained decoding and one-pass option scoring of one Qwen3-0.6B checkpoint, supervised references including learned decision heads, and a commercial System One model with identical states, answer spaces and outcomes. First results show that validity and information separate. Free and constrained generation return a valid distribution on every attempt, yet reproduce the prompt's uniform format example on all 15,491 outputs per readout; one-pass scoring never does. The hosted System One model is schema-valid on every agent-trace request, but its test log loss (0.7986 nats raw, 0.7747 after temperature scaling) exceeds that of a uniform prediction (0.6931); on maritime and weather tasks it assigns zero probability to eleven observed calibration events per family, leaving temperature scaling without a finite-loss solution. Machine-native decisions must be evaluated as probabilities on outcomes, not as outputs that fit a schema.

1 INTRODUCTION

Machine intelligence is increasingly embedded in software that consumes a model's output as a typed value rather than as text: a choice among available tools or routes, the probability that an event will occur, or a grade on a fixed scale. We call such decisions *machine-native*, because their consumer is a program that parses a value over a declared answer space and acts on it without reading an explanation. System One interfaces serve this use directly. Given a state and a typed question, they return a probability distribution over the supplied events instead of generating an answer token by token, and hosted System One models now expose this contract (TypeSafe AI, 2026). We use "System One" as a label for this interface only; it carries no claim about human cognition or about the internal mechanism of any evaluated model.

A typed interface guarantees form. Every answer names an admissible event and every vector sums to one, so the malformed outputs that burden text-based pipelines cannot occur. Form, however,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

is not information. A well-formed distribution can ignore the state, assign high confidence to the wrong event or assign zero probability to what actually happens. Whether a machine-native decision is useful therefore depends on properties that the interface cannot certify: accuracy and calibration on observed outcomes, reliability under temporal and entity shift, stability when the answer space is reordered or reworded, and the computation needed to deliver it. Calibration, selective prediction and shift have each been studied extensively (Guo et al., 2017; Geifman & El-Yaniv, 2019; Ovadia et al., 2019), and broad evaluations report them alongside accuracy (Liang et al., 2023). Forecasting benchmarks score probabilities against realized outcomes (Karger et al., 2025; Aksu et al., 2024; Williams et al., 2025). What the field lacks is a common test in which interchangeable decision interfaces return full distributions over the same typed, exhaustive event partitions and are scored on the same subsequently observed outcomes.

The gap is not hypothetical. We elicited probability vectors from Qwen3-0.6B by free generation and by constrained generation that admits only valid vectors. Without a format example in the prompt, generation exhausted its output budget on every tuning input of a formatting pilot; with one, it completed. Across all seven data families of our benchmark, both readouts then return a valid distribution on every attempt, and on all 15,491 outputs per readout that distribution is the prompt's uniform format example, unchanged by the input state. One-pass option scoring of the same checkpoint, which reads the next-token scores of the option labels, never reproduces it. The hosted System One model shows the same separation in a different form. On agent-trace outcomes it returns a valid distribution for every request, yet its test log loss is 0.7986 nats before and 0.7747 nats after temperature scaling, both above the 0.6931 nats of a uniform binary prediction. On the maritime and weather tasks it assigns probability zero to eleven observed calibration events in each family, so temperature scaling has no finite-loss solution. Schema validity, dependence on the state, accuracy and calibration are distinct properties, and each must be measured.

SystemOneBench measures them. Its unit of evaluation combines an observed state, a prediction cutoff, a typed question, an answer space built from pre-outcome information and a target set by the subsequently observed outcome. Seven data families span six domains: Coinbase limit-order books and IEX quotes in finance, Bitcoin mempool inclusion, highD lane changes (Krajewski et al., 2018), Daymet precipitation (Thornton et al., 2021), AIS vessel speed changes and τ^2-bench agent traces (Barres et al., 2025). Every implementation must return a complete distribution; invalid attempts count as abstentions in every coverage denominator, and exact zeros are scored rather than clipped. Forward-time, recording, location, identifier and owner-group holdouts, together with option permutation, paraphrase and held-out partitions, test whether probability quality survives changes in time, entity and answer space.

The benchmark turns the central claims about machine-native decision models into falsifiable tests. Whether bounded decisions need autoregressive generation becomes a paired contrast between readouts of one checkpoint on identical outcomes. Whether a fast decision layer knows when to defer to slower reasoning becomes a question of calibration and selective risk under frozen thresholds. Whether a decision function generalizes becomes a question of probability quality across families, shifts and answer spaces. The same fixed outcomes, partitions and cost accounting apply to later readouts, distilled decision models and confidence-based routers. Our comparisons concern the evaluated implementations: three readouts of one pinned Qwen3-0.6B checkpoint with one anchored generation prompt, and a hosted service of undisclosed architecture and revision. They do not rank model families or architectures.

We make four contributions.

- A protocol for typed probabilistic decisions: exhaustive event partitions constructed from pre-cutoff information, outcome-derived labels, a probability contract that keeps invalid attempts in all coverage denominators and retains exact zeros, and paired analyses of calibration, selective risk, shift and answer-space perturbation with dependence-aware resampling.
- An instantiation across seven data families in six domains, with forward-time and entity holdouts and declared dependence units for every family.
- A matched evaluation harness in which free-generation, constrained-generation and one-pass readouts of one checkpoint, a hosted System One model and supervised references

receive identical states, answer spaces and outcomes under fixed selection and calibration rules.

- Empirical findings from 46,473 local attempts and 15,408 hosted requests: generated distributions that are valid on every attempt yet reproduce a format example on every output, and hosted distributions that are valid but less informative than a uniform prediction or that assign zero probability to observed events.

2 Related work

Eliciting probabilities from language models. Kadavath et al. (2022) study self-evaluation through $P(\mathrm{True})$ and a fine-tuned $P(\mathrm{IK})$ head whose calibration degrades on new tasks. Tian et al. (2023) compare verbalized confidence with conditional answer probabilities, and Xiong et al. (2024) decompose black-box elicitation into prompting, sampling and aggregation, finding that no strategy dominates. Gruver et al. (2023) convert token distributions over serialized digits into forecast densities. Single-token option scoring builds on multiple-choice prompting and symbol binding (Robinson et al., 2023); option-order sensitivity is established (Pezeshkpour & Hruschka, 2024), and first-token choices can disagree with generated answers (Wang et al., 2024). These findings motivate treating each readout as a distinct decision implementation. Our contribution is not a new elicitation method but a protocol in which every implementation returns a full distribution over a supplied exhaustive partition of future events and is scored on shared observed outcomes.

Calibration, selective prediction and shift. Temperature scaling is the standard post-hoc calibration baseline (Guo et al., 2017); SelectiveNet learns prediction and rejection jointly (Geifman & El-Yaniv, 2019), whereas we evaluate a post-hoc confidence threshold. HELM reports calibration and robustness alongside accuracy (Liang et al., 2023). Ovadia et al. (2019) show that uncertainty quality degrades under dataset shift and that temperature scaling fitted in distribution loses its advantage as shift grows; Wild-Time benchmarks naturally occurring temporal shift with fixed and streaming splits (Yao et al., 2022). Measuring these properties together is therefore not novel by itself; we measure them for typed decisions whose event partitions come from pre-outcome state.

Outcome-scored forecasting. ForecastBench scores forecasts of questions unresolved at submission (Karger et al., 2025), protecting against contamination in a way our retrospective splits cannot. FinBench requires time-gated structured probabilities of realized returns scored by Brier and interval scores (Ghosh & Devarakonda, 2026). GIFT-Eval (Aksu et al., 2024) and Context is Key (Williams et al., 2025) evaluate probabilistic forecasters across seven domains with CRPS, the latter relating quality to model size and inference time. Paleka et al. separate consistency across related questions from Brier scores on resolved outcomes (Paleka et al., 2024), as our permutation and paraphrase diagnostics do. Domain benchmarks exist for limit-order books (Ntakaris et al., 2018), probabilistic weather forecasting (Garg et al., 2022), calibrated lane-change anticipation on highD (Kılınç et al., 2026) and early prediction of agent success from partial trajectories (Shi et al., 2026); their data and targets differ from ours, so their scores are not comparable. SystemOneBench narrows these precedents to one question: given a finite, typed and exhaustive set of outcome events defined from pre-cutoff state, do interchangeable decision interfaces return distributions that are valid, state-dependent, accurate and calibrated on the same observed outcomes?

3 Benchmark protocol

Prediction unit. Let X_t be the state available at a declared cutoff t and $Y_{t,h}$ an outcome observed over horizon h. A question Q identifies the target variable, and an answer space $A = (E_1, \ldots, E_K)$ partitions its support into mutually exclusive, exhaustive events. A model returns

$$p_k(X_t, Q, A) \approx \Pr(Y_{t,h} \in E_k \mid X_t, Q, A), \qquad p_k \geq 0, \quad \textstyle\sum_{k=1}^{K} p_k = 1.$$

Each event has a stable semantic identifier distinct from its display position; numerical intervals carry units and endpoint conventions, and an other-outcome cell is added when named cells would omit valid outcomes. The answer space is built from the task definition and cutoff-known information before the target is attached, so future values and outcome-derived identifiers cannot influence option selection. Scores from different partitions are reported separately, never pooled.

Labels and prediction views. A deterministic task rule maps each subsequently observed source outcome to exactly one event; targets are never generated reference answers. Source adapters distinguish labeled examples from incomplete histories, censored windows and excluded data, because an absent observation does not show that an event failed to occur. The prediction view (state, question, events) is separate from the evaluation view (outcome and allocation metadata), and hosted requests carry only the prediction view. Repeated displays of one outcome stay paired and inherit its split.

Probability contract. Every implementation must return one finite, nonnegative value per event identifier, summing to one within its declared tolerance. Parsing failures, incomplete vectors, exhausted output budgets, over-length inputs and failed hosted requests are invalid attempts. Completed invalid responses are neither repaired nor regenerated; they remain in the all-attempt denominator of every coverage statistic and abstain in selective prediction. Exact zeros are retained: a zero on the realized event gives infinite log loss, so every log-loss summary reports the number of such outcomes alongside finite Brier and ranked probability scores, and no probability floor is imposed.

Split order. Training data fit transforms, thresholds and model parameters; tuning data select hyperparameters, prompts and comparison representatives; calibration data fit only temperatures and selective thresholds; test outcomes are scored with everything frozen. Temporal families follow forward calendar order with purging of windows that cross a partition boundary. highD holds out whole recordings and a separate location, and Agent holds out owner components (Section 4).

Scores, calibration and selective prediction. Negative log-likelihood $-\log p_{i,y_i}$ and the multiclass Brier score $\sum_k (p_{ik} - \mathbf{1}\{y_i = k\})^2$ are co-primary; the ranked probability score applies to declared ordinal targets, and top-label calibration error uses 15 fixed equal-width bins. Temperature scaling uses power normalization, $p_k^{(T)} \propto p_k^{1/T}$ with $T \in [0.2, 5]$ fitted by calibration log loss. Because this map preserves zeros, a valid calibration response that assigns zero to its observed event makes calibration loss infinite at every temperature; we then retain $T = 1$ and flag that no finite-loss solution exists. Selective prediction thresholds the largest event probability at calibration acceptance targets of 0.5, 0.8 and 0.95 of all attempts, and frozen thresholds are compared on test with random abstention at the same achieved valid coverage.

Answer-space changes and shift. Permutations reorder the same semantic events and a fixed paraphrase rewords them; held-out partitions, where a task catalog defines them, change the events themselves. Aligned by event identity, pairs are compared by total variation $\frac{1}{2}\sum_k |p_k - p'_k|$, top-prediction changes and paired loss changes. Forward-time strata apply to LOB, IEX, mempool, weather and AIS; the highD recording and location tests, the weather spatial holdout, the AIS identifier holdout and the Agent owner-component holdout are entity or location shifts.

Paired contrasts and uncertainty. The protocol fixes four primary contrasts on common valid outcomes: (i) one-pass scoring versus free generation of the same checkpoint; (ii) raw versus temperature-scaled outputs; (iii) the learned head versus the tuning-selected domain control; and (iv) the hosted model versus the tuning-selected local readout. Every local contrast is reported beside a uniform distribution over the displayed options ($\log K$ nats), together with the fraction of outputs equal to the prompt's format example and a state-dependence check, so that a readout that ignores the state cannot pass as a representative unnoticed. Intervals come from 1,000 dependence-aware resamples whose unit or block multiplicities are shared across methods, calibration variants, option variants and metrics (Table 3; Appendix C). Axes with fewer than ten test units or five calibration units support descriptive diagnostics only. Appendix D specifies local CPU latency, throughput and package-energy measurement on a fixed per-family workload, separately from client-observed hosted latency.

4 SEVEN-FAMILY INSTANTIATION

Table 1 summarizes the seven families; Appendix A gives their construction and Table 2 their eligible cohort sizes. The families share the protocol but differ in horizon, arity, event prevalence, dependence structure and source boundary, which is the point: a decision interface is tested on con-

Family (domain)	Prediction target	Test allocation and extra holdout	Dependence unit
LOB (finance)	1-s Coinbase BTC-USD midpoint return from 11 reconstructed book states	Forward sessions; 3 test sessions	Calendar session
IEX (finance)	5-s midpoint return from 30 s of displayed quotes	Forward trading days; 2 test days	Trading day
Mempool (blockchain)	Time to block inclusion within 1,800 s, or non-inclusion	Forward collector day; 1 test day	Collector day
highD (traffic)	Signed lane displacement 2 s ahead	Held-out recordings; held-out location (3 recordings)	Recording
Weather (weather)	Next product-day precipitation interval, Puerto Rico lattice	Test years 2020–2025; 36 held-out cells	Date within test year
AIS (maritime)	Reported speed-change interval over 10 min	Forward dates (January 2022); 14 held-out identifiers	Date $\times$ identifier
Agent (τ^2-bench)	Terminal database agreement under replay	Held-out owner components; 11 test components	Owner component

Table 1: The seven SystemOneBench families. Targets are computed from subsequently observed source outcomes; dependence units define the paired resampling of Table 3.

Family	Training	Tuning	Calibration	Test	Additional holdout
LOB	14,268	2,378	1,189	3,567	—
IEX	540	270	270	540	—
Mempool	1,604	549	85	1,417	—
highD	639,313	159,748	147,288	173,812	27,007
Weather	2,405,760	300,720	300,720	257,760	77,328
AIS	87,763	16,174	17,574	12,925	7,327
Agent	1,072	224	272	288	—

Table 2: Eligible outcomes in the constructed cohorts before model sampling caps. Additional holdouts contain highD location recordings, weather spatial cells and outcomes of held-out AIS identifiers. Counts describe eligible observations, not independent samples.

tinuous market microstructure, discrete inclusion times, physical interaction, geophysical products, noisy self-reported motion and symbolic software state.

Finance. The LOB task reconstructs Coinbase level-3 books from snapshots and contiguous full-channel sequences with exact integer price arithmetic; the conditional cohort retains 18 of 65 allocated sessions (21,402 eligible outcomes) after integrity and continuity checks. IEX is an opening-window pilot on AAPL, SIRI and GEVO over six capture dates, with 540 training, 270 tuning, 270 calibration and 540 test outcomes.

Mempool. Inclusion is recognized only after both a block's transaction list and its connect marker appear, with six-deep closure support. One collector day is split at fixed UTC boundaries into training, tuning and calibration, and a later collector day is the test set (1,604, 549, 85 and 1,417 transactions).

Traffic, weather and maritime. highD examples require complete two-second histories and outcomes; velocities are recomputed from past positions and 1,127,959 of 1,147,168 eligible cutoffs have no lane change, an imbalance preserved in scoring. The weather task uses the preceding seven Daymet V4 product days on 180 cells, divided into development, buffer and held-out cells. AIS uses three past reports of a fixed identifier sample over 215 days; 78,281 of 141,763 eligible outcomes are exact zero speed changes, so a no-change control is essential.

Agent traces. The state ends immediately before the first assistant tool result of a τ^2-bench episode and excludes later text, rewards, reference actions and hidden simulator instructions. The target is whether the reconstructed terminal database equals the reference database under a pinned

Family	Dependence unit	Resampling of the test stratum
LOB	Calendar session	Noncircular moving blocks of three sessions (one and seven as sensitivities) over the 13-position test calendar; empty positions retained
IEX	Trading day; symbols together	Empirical resampling of the two test days
Mempool	Collector day; transaction identities together	None: single test day; paired point estimates only
highD	Recording; interacting vehicles together	Recording-level resampling, separately for recording and location tests
Weather	Date within test year; all cells of a date together	Noncircular 30-day blocks (7 and 60 as sensitivities) within each test year; development and held-out cells separately
AIS	Date crossed with identifier	Noncircular seven-day blocks over the 31-day January test calendar, crossed with identifier resampling
Agent	Owner-connected component	Empirical resampling of the eleven test components

Table 3: Dependence units and paired resampling schemes; 1,000 shared draws per family and stratum.

runtime. Of 2,624 episodes, 1,856 across 116 tasks are labeled (1,241 matches, 615 mismatches). Owner components keep all agents and trials of a task together; the test set contains 288 episodes from 18 tasks in eleven components.

5 Evaluated implementations

Three readouts of one checkpoint. All local readouts use Qwen/Qwen3-0.6B at a pinned revision in float32 on the CPU with greedy decoding. *One-pass* scoring applies a softmax to the next-token logits of the single-token option letters in one forward pass. *Free* and *constrained* generation request one nonnegative integer weight per option summing to 1,000 within $6K + 8$ output tokens for K options; the constrained decoder admits only prefixes that can complete a valid vector, and both require a terminal end-of-sequence token. Inputs longer than 4,096 tokens are rejected without truncation. A formatting-only pilot on twelve tuning inputs compared a plain prompt with one containing a cardinality-matched, approximately uniform weight vector as a format example: the plain prompt exhausted its output budget on all twelve inputs, and all twelve responses to the example prompt copied the example. The example prompt was fixed for evaluation.

Hosted System One model. The Jev System One model (TypeSafe AI, 2026) is queried through its hosted evaluation interface with the same state, question and typed event descriptions. Options receive positional letter keys and returned probabilities are mapped back to event identities; a response must cover every option, lie in $[0, 1]$ and sum to one within 10^{-8}, and exact zeros stay zero. The hosted allocation covers six families; highD trajectories stay local under their source-use terms. The service architecture, size and revision behind its alias are not disclosed, so differences from the local model are not attributed to architecture. A recovery protocol for rate-limited requests (Appendix B) resends only recognized transport or transient failures and keeps every completed response as returned.

Supervised references. Base-rate or climatology predictors, persistence, regularized linear models, boosted trees and a per-task learned event scorer without pretrained weights are fitted on the training partitions (Appendix B). Among them, this paper reports the Agent base-rate control.

Allocations. Within each split a fixed seed selects examples uniformly before labels are inspected, with caps of 1,024 tuning, 2,048 calibration and 1,024 test examples per stratum and 256 test outcomes per paired option variant. For highD, AIS and weather the local readouts use smaller nested allocations chosen without consulting outcomes to cover the dependence groups: 128 tuning, 256 calibration and 256 outcomes per test stratum, with 32 paired variants per test split (Appendix Table 6).

Family	Domain	Attempts per readout	Valid (%, all readouts)	Equal to format example (% of valid)		
				Free	Constrained	One-pass
LOB	Finance	4,773	100.0	100.0	100.0	0.0
IEX	Finance	2,132	100.0	100.0	100.0	0.0
Mempool	Blockchain	3,194	100.0	100.0	100.0	0.0
highD	Traffic	1,024	100.0	100.0	100.0	0.0
Weather	Weather	1,536	100.0	100.0	100.0	0.0
AIS	Maritime	1,536	100.0	100.0	100.0	0.0
Agent	Software agents	1,296	100.0	100.0	100.0	0.0
All	Six domains	15,491	100.0	100.0	100.0	0.0

Table 4: Output validity and format-example reproduction for three readouts of Qwen3-0.6B. Counts include tuning, calibration, test and option-variant displays; each readout receives the same inputs. "Equal to format example" is an exact comparison of the returned probability vector with the prompt's cardinality-matched example.

6 RESULTS

6.1 VALIDITY IS NOT INFORMATION

Table 4 reports, for every family and local readout, the attempted decisions, the valid fraction and the fraction of valid outputs that equal the prompt's format example exactly. All 46,473 local attempts return a valid distribution. Free and constrained generation reproduce the format example on every one of their 15,491 outputs per readout, across tuning, calibration, all test strata and the permutation, paraphrase and held-out-partition displays (Appendix Table 6). One-pass scoring reproduces it on none.

Because the example depends only on the number of displayed options, the free and constrained outputs are a fixed function of the option count and display order: they are identical to each other on every paired input and are unchanged by the state. Their proper scores therefore equal those of the uniform reference $p_k = 1/K$ up to the rounding of the example to multiples of 0.001, which is nonzero for three- and six-option spaces. Temperature scaling can amplify these rounding differences, but any resulting score change carries no information about the state, which is why the protocol reports the uniform reference and format-example fraction beside every local contrast. The constrained decoder thus does exactly what it promises, and that promise is not the one a decision consumer needs. One-pass scores never copy the example, but not copying is not the same as being informative; that question is answered only by proper scores against outcomes.

6.2 A VALID HOSTED DECISION CAN BE WORSE THAN UNIFORM

On Agent traces, the hosted model returns a valid distribution for all 1,072 logical requests: 272 calibration episodes, 288 base test episodes and 256 paired displays each for permutation and paraphrase. Table 5 (left) reports its base-test scores. Temperature scaling fitted on the calibration episodes ($T = 1.1745$) reduces test log loss by 0.0239 nats, yet the calibrated loss remains 0.0816 nats above the uniform binary reference, and the raw loss 0.1055 nats above it. Complete schema validity thus coexists with outcome error larger than that of a prediction that ignores the state. The base-rate control is also worse than uniform. A constant prediction whose rate equals the test frequency has log loss at most $\ln 2$, so a constant control at 0.8408 nats shows that the training frequency of database matches differs substantially from the test frequency: the owner-group holdout shifts the base rate itself. These scores describe 288 episodes from 18 tasks in eleven owner components, not 288 independent tasks.

6.3 EXACT ZEROS BREAK CALIBRATION

On AIS and weather the recovered hosted runs return 7,109 (99.2%) and 7,084 (98.8%) valid responses of 7,168 logical requests; the remaining 59 and 84 violate the probability or response contract and stay in the coverage denominators (Table 5, right). No unresolved transport failure

Agent base test ($n = 288$)	Acc. (%)	NLL
System One, raw	48.26	0.7986
System One, $T = 1.1745$	—	0.7747
Base-rate control	47.22	0.8408
Uniform reference	—	0.6931

Hosted support	AIS	Weather
Valid / requests	7,109/7,168	7,084/7,168
Valid calibration	2,036/2,048	2,023/2,048
Zeros on observed event	11	11
Temperature	1 (fallback)	1 (fallback)

Table 5: Hosted System One model. Left: Agent base-test accuracy and log loss (nats) with the uniform binary reference ($\ln 2$) and the base-rate control; temperature scaling leaves top-label accuracy unchanged. Right: output support on AIS and weather, with zeros counted among valid calibration responses; allocations include repeated option displays of test outcomes, and invalid responses remain in the denominators.

remains. Each calibration set contains eleven valid distributions that assign exactly zero probability to the event that occurred. Because power normalization preserves zeros, calibration log loss is infinite at every admissible temperature, and both families fall back to $T = 1$. Several ordinary and shifted test partitions likewise have infinite log loss despite finite Brier and ranked probability scores. A clipping floor would hide this failure; reporting exact-zero counts keeps it visible. For a consumer that acts on the probability, a zero is a claim of impossibility, and here that claim is falsified on roughly one in two hundred calibration outcomes.

7 Discussion

In these runs, constraining the output format produced well-formed outputs, not decisions. A decision interface is useful only if its probabilities depend on the state and track outcome frequencies; neither property follows from the output format, and both have to be measured. The case for machine-native decision models rests on four empirical questions, and the benchmark reduces each to a measurement.

Do bounded decisions need generation? The paired readouts of one checkpoint are the right instrument, but the free and constrained arms of this comparison produced a template. Any contrast between generative and one-pass readouts must therefore report the uniform reference and a state-dependence check, or it risks comparing a decision readout with the prompt. *Can a reasoning model be compiled into a fast decision model?* Distillation replaces generated answers with a student's distribution over the same events. Because the benchmark never uses generated or teacher answers as targets, a distilled student is scored on observed outcomes, and its agreement with the teacher can be reported separately from its calibration on those outcomes. *Can a fast layer know when to escalate?* Confidence routing from a System One layer to a slower deliberative system needs finite, calibrated probabilities. Exact zeros on observed events make calibration log loss infinite, disable temperature fitting and undermine any log-loss-based selection of a routing threshold; a zero on an event that then occurs is an error the model has reported as impossible. *Does a cross-domain decision function exist?* Any such claim has to hold on outcome-labelled families under temporal and entity shift. The Agent base-rate control already shows that an owner-group holdout moves the base rate far enough that a constant prediction trails the uniform one; a general decision function has to survive shifts of this kind in every family.

Independent evaluation of hosted decision models. The published evaluation of the hosted model we test compares its probabilities with reference probabilities produced by other models (TypeSafe AI, 2026). Agreement with a model reference and calibration on outcomes answer different questions; a consumer that acts on the probability relies on the second, which can be measured only against outcomes that neither the provider nor a reference model produces. Our hosted results are one such check, specific to one service alias, the reported families and the requests we sent; they do not rank the service against other models, and they do not show what a reference-based evaluation would have missed.

The benchmark supplies what these tests require: fixed outcomes and answer spaces, all-attempt denominators, retained zeros, declared dependence units and a uniform reference, applied identically to local readouts, hosted services, supervised references and future distilled decision models or routers.

8 LIMITATIONS

Our findings concern one pinned Qwen3-0.6B checkpoint, one generation prompt anchored by a uniform format example, and a hosted service of undisclosed architecture whose alias does not identify a fixed revision; they do not characterize model families, scales or architectures. The sources bound interpretation: Daymet is a retrospective product rather than an operational forecast, highD trajectories retain the vendor's smoothing over all detections, the LOB cohort uses exchange time and a conditional subset of sessions, and the Agent target is database equality under replay rather than complete task success, with owner components that share catalog items. Some axes have few units (three LOB test sessions, two IEX test days, three highD location recordings, one held-out weather region), and the mempool test split covers a single collector day, so we do not estimate day-level uncertainty for this family. The reduced highD, AIS and weather local cohorts cover dependence groups and do not estimate performance at natural prevalence. The analysis rules were fixed after test outcomes for several families had been inspected and are not preregistered, and the seven families are not a representative sample of machine-consumed decisions.

Scope of the reported results. This paper reports validity, format-reproduction and hosted-support diagnostics for all seven families and outcome scores of the hosted model on agent traces. The protocol's paired proper-score, calibration, selective-risk, shift and option-perturbation contrasts for every family and implementation, and the local cost measurements of Appendix D, apply unchanged to the saved predictions but are not reported here; neither are other checkpoints, scales or prompts without a copyable example.

9 CONCLUSION

SystemOneBench evaluates software-consumed predictions as typed probability distributions scored against subsequently observed outcomes, across seven data families in six domains, and applies one probability contract to generative readouts, one-pass scorers, supervised references and hosted System One models. Its first results show that output validity and information come apart. Free and constrained generation from Qwen3-0.6B satisfy the output contract on every attempt while reproducing the prompt's format example on every output; one-pass scoring of the same checkpoint never does. The hosted System One model is schema-valid on agent traces yet scores worse than a uniform prediction, and it assigns zero probability to observed maritime and weather events, leaving zero-preserving temperature scaling without a finite-loss solution. These findings concern the evaluated implementations, not model families or architectures; their implication is general. A machine-native decision is only as useful as the probability behind it, and that probability has to be measured on outcomes, independently of the interface that returns it. The benchmark's fixed outcomes, answer spaces and accounting rules provide common ground for the next questions in this area: whether one-pass decision architectures deliver more information per unit of computation than generation, whether reasoning models can be distilled into calibrated decision models, and when a fast decision layer should defer to deliberative reasoning.

STATEMENTS

REPRODUCIBILITY STATEMENT

Section 3 and Appendix C define the probability contract, split order, scoring, calibration, selective-prediction and resampling rules; Appendix A and Table 2 specify every family's construction and cohort; Appendix B gives the readouts, the supervised references and the hosted interface; and Appendix E lists the checkpoint revision, seeds and sources. Reproduction metadata include configurations, input versions, seeds, thread settings, software environments, and content digests for registered inputs. Local inference is deterministic greedy decoding, run once per configuration. Hosted results depend on a service whose revision is not disclosed and may not be exactly reproducible. Code and instructions are provided in the supplementary material.

AI USE STATEMENT

Generative AI tools were used to assist with retrieving and discovering related literature, code implementation, writing and polishing. The research ideas originated with the authors. The authors reviewed and verified AI-assisted content and take full responsibility for the paper's accuracy, originality, and conclusions.

REFERENCES

Taha Aksu, Gerald Woo, Juncheng Liu, Xu Liu, Chenghao Liu, Silvio Savarese, Caiming Xiong, and Doyen Sahoo. Gift-eval: A benchmark for general time series forecasting model evaluation. arXiv preprint arXiv:2410.10393, 2024. URL https://arxiv.org/abs/2410.10393.

Victor Barres, Honghua Dong, Soham Ray, Xujie Si, and Karthik Narasimhan. τ^2-bench: Evaluating conversational agents in a dual-control environment, 2025. URL https://arxiv.org/abs/2506.07982.

Sagar Garg, Stephan Rasp, and Nils Thuerey. Weatherbench probability: A benchmark dataset for probabilistic medium-range weather forecasting along with deep learning baseline models. *arXiv preprint*, 2022. URL https://arxiv.org/abs/2205.00865v1.

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2151–2159. PMLR, 2019. URL https://proceedings.mlr.press/v97/geifman19a.html.

Rishab Ghosh and Vinay Devarakonda. FinBench: Time-gated calibration and uncertainty benchmarking for agentic financial forecasting, 2026. URL https://arxiv.org/abs/2607.16229. arXiv preprint.

Nate Gruver, Marc Finzi, Shikai Qiu, and Andrew Gordon Wilson. Large language models are zero-shot time series forecasters. In *Advances in Neural Information Processing Systems 36*, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/3eb7ca52e8207697361b2c0fb3926511-Abstract-Conference.html.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know. arXiv preprint arXiv:2207.05221, 2022. URL https://arxiv.org/abs/2207.05221.

Ezra Karger, Houtan Bastani, Chen Yueh-Han, Zachary Jacobs, Danny Halawi, Fred Zhang, and Philip E. Tetlock. ForecastBench: A dynamic benchmark of AI forecasting capabilities. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/ea74e45a229dac70b5b63b28d8934db6-Paper-Conference.pdf.

Dilara Kılınç, Hendrik Kleikamp, and Bruno Viti. Robust lane-change intention anticipation under uncertainty based on a recursive bayesian filtering approach. *arXiv preprint*, 2026. URL https://arxiv.org/abs/2608.21132v1.

Robert Krajewski, Julian Bock, Laurent Kloeker, and Lutz Eckstein. The highD dataset: A drone dataset of naturalistic vehicle trajectories on german highways for validation of highly automated driving systems. In *IEEE 21st International Conference on Intelligent Transportation Systems (ITSC)*, 2018. URL https://arxiv.org/abs/1810.05642.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023. URL https://arxiv.org/abs/2211.09110.

Adamantios Ntakaris, Martin Magris, Juho Kanniainen, Moncef Gabbouj, and Alexandros Iosifidis. Benchmark dataset for mid-price forecasting of limit order book data with machine learning methods. *Journal of Forecasting*, 37(8):852–866, 2018. doi: 10.1002/for.2543. URL https://arxiv.org/abs/1705.03233v5.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D Sculley, Sebastian Nowozin, Joshua V. Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/8558cb408c1d76621371888657d2eb1d-Paper.pdf.

Daniel Paleka, Abhimanyu Pallavi Sudhir, Alejandro Alvarez, Vineeth Bhat, Adam Shen, Evan Wang, and Florian Tramèr. Consistency checks for language model forecasters. *arXiv preprint*, 2024. URL https://arxiv.org/abs/2412.18544.

Pouya Pezeshkpour and Estevam Hruschka. Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 2006–2017. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-naacl.130. URL https://aclanthology.org/2024.findings-naacl.130/.

Joshua Robinson, Christopher Michael Rytting, and David Wingate. Leveraging large language models for multiple choice question answering. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.12353v3. Author manuscript, arXiv:2210.12353v3.

Yuling Shi, Zhensu Sun, Junsen Dong, Chengcheng Wan, David Lo, and Xiaodong Gu. Earlyeval: Cheaper agent evaluation via early outcome prediction. *arXiv preprint*, 2026. URL https://arxiv.org/abs/2609.02783v1.

Peter E. Thornton, Rupesh Shrestha, Michele Thornton, Shih-Chieh Kao, Yaxing Wei, and Bruce E. Wilson. Gridded daily weather data for north america with comprehensive uncertainty quantification. *Scientific Data*, 8:190, 2021. doi: 10.1038/s41597-021-00973-0. URL https://www.nature.com/articles/s41597-021-00973-0.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D. Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. doi: 10.18653/v1/2023.emnlp-main.330. URL https://aclanthology.org/2023.emnlp-main.330/.

TypeSafe AI. Introducing System One Models & Jev. https://typesafe.ai/blog/introducing-system-one-models-and-jev, 2026. Accessed September 2026.

Xinpeng Wang, Bolei Ma, Chengzhi Hu, Leon Weber-Genzel, Paul Röttger, Frauke Kreuter, Dirk Hovy, and Barbara Plank. "my answer is C": First-token probabilities do not match text answers in instruction-tuned language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 7407–7416. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.441. URL https://aclanthology.org/2024.findings-acl.441/.

Andrew Robert Williams, Arjun Ashok, Étienne Marcotte, Valentina Zantedeschi, Jithendaraa Subramanian, Roland Riachi, James Requeima, Alexandre Lacoste, Irina Rish, Nicolas Chapados, and Alexandre Drouin. Context is key: A benchmark for forecasting with essential textual information. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 66887–66944, 2025. URL https://proceedings.mlr.press/v267/williams25a.html.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2306.13063.

Huaxiu Yao, Caroline Choi, Bochuan Cao, Yoonho Lee, Pang Wei Koh, and Chelsea Finn. Wild-Time: A benchmark of in-the-wild distribution shift over time. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/43119db5d59f07cc08fca7ba6820179a-Paper-Datasets_and_Benchmarks.pdf. Datasets and Benchmarks Track.

A FAMILY CONSTRUCTION

Table 2 in Section 4 gives the eligible cohort of each family.

LOB. The task predicts the one-second midpoint return from eleven reconstructed Coinbase BTC-USD book states spanning the preceding ten seconds. Level-3 snapshots initialize the book; contiguous full-channel sequence numbers update resting orders, including maker fills, cancellations and price modifications. One-second grids include all events at the grid timestamp and use exact integer price arithmetic. The state contains best prices, BTC quantities, top-25 aggregated depth, observed level counts and time since the last exchange event. Receipt timestamps are unavailable, so this is retrospective exchange-time forecasting. The cohort selects the earliest stored snapshot on each of 65 session dates and a 20-minute interval after alignment. Chronological allocations contain 39 training, six tuning, seven calibration and 13 test sessions; the test sessions span November 11–23, 2022. Forty-seven sessions fail snapshot integrity, event continuity, book consistency or interval completion checks, leaving 18 sessions (12 training, two tuning, one calibration and three test) and 21,402 eligible outcomes. This conditional cohort does not represent complete trading days, and exclusions may depend on market activity.

IEX. Six local capture files are allocated chronologically to two training dates, one tuning date, one calibration date and two test dates; two further date directories contain no capture file. Each source is replayed from its beginning, and one-second origins are retained during the first two minutes after the regular-session start message for the prespecified symbols AAPL, SIRI and GEVO. A ten-minute extraction failed on five truncated captures, so the common two-minute window was fixed before comparative model evaluation. Each input contains thirty seconds of displayed bid and ask history, and the target is the midpoint return five seconds later. Times use the recorded capture clock without retrospective correction.

Mempool. The cutoff is completion of a transaction-add log entry after enrichment. Features contain the recorded transaction attributes, counts and volume-weighted fee statistics from the preceding 600 seconds, and the age of the latest fully observed block. The target is elapsed collector time to inclusion within 1,800 seconds, with a separate non-inclusion outcome. Every labeled window requires common continuity and six-deep closure support. The first collector day (August 12, 2026) is allocated at fixed UTC boundaries: before noon to training, noon to 18:00 to tuning, and the remaining hours to calibration; September 18, 2026 is the test day. Windows whose feature and outcome support cross a boundary are purged (350 windows), and repeated transaction identities cannot span splits. The calibration observations concentrate within approximately 53 minutes.

highD. The task built from highD (Krajewski et al., 2018) predicts signed lane displacement two seconds after a one-second-grid cutoff. Features describe the ego vehicle and six nearest longitudinal neighbor slots at three past frames. Vendor velocity and trajectory-summary fields are excluded, but the source trajectories retain the vendor's RTS smoothing, so the task does not establish online sensor causality. Whole recordings define the split unit under a fixed metadata-based allocation, and three recordings from a held-out location form a separate test; recording identifiers do not establish calendar order. Of the 1,147,168 eligible cutoffs, 1,127,959 have no endpoint lane displacement, 8,702 move one lane left and 10,507 one lane right.

Weather. The implementation uses Daymet V4 (Thornton et al., 2021) on a Puerto Rico lattice of 180 cells: 120 development, 24 buffer and 36 held-out cells. Training covers 1950–2005, tuning 2006–2012, calibration 2013–2019 and forward testing 2020–2025. Each prediction uses the preceding seven product days and a target day in the same annual file, with Daymet's leap-year convention handled explicitly. This is retrospective prediction from a processed observation product, not an evaluation of operational availability.

AIS. The task predicts reported speed change over a ten-minute grid horizon from three past observations. A fixed identifier sample spans all 215 days from July 2021 through January 2022. The entity test comprises the 7,327 outcomes of fourteen reported identifiers excluded from every fitting period that have eligible outcomes on the forward-test dates; these are identifiers as reported, not verified physical vessels. Each grid endpoint permits an observation age of up to sixty seconds.

Agent. The source environment is τ^2-bench (Barres et al., 2025). The prediction state ends immediately before the first assistant tool result and includes the preceding tool request. Of 2,624 episodes, 1,904 have a changed, error-free reference comparison; excluding 48 episodes attached to known defective references leaves 1,856 labeled episodes. Allocations keep all source-agent and trial observations of each task within audited owner/order/reservation components (23 airline and 45 retail). Within each domain, a seeded shuffle of sorted components determines training, tuning, calibration and test sets of 1,072, 224, 272 and 288 episodes (67, 14, 17 and 18 tasks) without consulting labels. Shared catalog items are permitted across splits, so this is an owner-group holdout, not an all-entity-disjoint evaluation; under one-hop catalog overlap, the 89 observed retail tasks form one connected component.

B Implementation details

Local readouts. The backbone runs in float32 on the CPU with scaled dot-product attention; its chat template disables the optional thinking mode. The payload is compact JSON with sorted object keys and preserved option order, and display letters replace semantic identifiers. Input length is measured after applying the chat template. The one-pass softmax is taken only over the displayed letters, each checked to occupy one tokenizer token, so its probabilities are conditional on the letter set rather than the full vocabulary. The generative readouts use greedy decoding with at most $6K+8$ new tokens. Constrained decoding permits only continuations whose decoded suffix can extend to a complete integer-simplex array. Invalid arrays and exhausted budgets remain failed decisions.

Supervised event scorer. State and event text projections use signed feature hashing of word unigrams and bigrams with ℓ_2 normalization, 1,024-dimensional for the state and question and 128-dimensional per event. Numeric values are transformed by $\operatorname{sign}(v)\log(1+|v|)$ and standardized on training states; absent fields receive a zero and an observation indicator. Event vectors include transformed interval bounds, bound-presence indicators, endpoint closure and category flags; identifiers and display positions are excluded, so predictions are permutation-equivariant. With 64-dimensional tanh encodings x and e_k,

$$s_k = w^\top \tanh(W[x; e_k; x \odot e_k] + b) + c, \qquad p_k = \exp(s_k) \big/ \textstyle\sum_j \exp(s_j).$$

Training minimizes cross-entropy with AdamW for 12 epochs in batches of 128. Seed zero selects among learning rates $\{0.001, 0.003\}$ and weight decays $\{0.0001, 0.01\}$ by tuning log loss; seeds one and two refit the selected setting.

Domain controls. Controls predict a distribution over the underlying outcome and integrate its mass over each supplied event. Categorical models use the task's fixed outcome vocabulary. Numerical models predict signed-log outcomes, estimate their distribution from tuning residuals and mix in a one-percent training-scaled prior component for support outside observed residuals. Mempool predictions factor inclusion probability and conditional elapsed time. Base-rate predictors preserve exact observed atoms, including zero changes and interval endpoints.

Hosted interface and recovery. Hosted requests use the Vercel AI Gateway through the `experimental_evaluate` interface of AI SDK 7.0.107 with the model alias `typesafe-ai/jev`. The hosted allocation fixes 23,664 logical requests across six families. The initial pass encountered extensive rate limiting with automatic retries disabled. The recovery protocol retains valid answers and completed invalid responses unchanged and resends only requests with recognized transport or transient service failures, with at most six recovery calls per logical request, a 30-second call timeout, nominal one-second spacing between call starts, provider-specified retry waits when available and exponential fallback waits otherwise. A required wait above 60 seconds, exhausted attempts or a non-retryable failure stops a recovery batch. All physical attempts are retained, and changing the service tier or retry policy does not create additional outcome samples.

C Analysis specification

Representatives and state dependence. Representatives for contrasts (iii) and (iv) are selected by mean raw log loss on tuning outcomes for which every candidate is valid, with ties following

the declared candidate order. When all candidates have infinite mean log loss, minimum mean Brier score decides, a fallback introduced after test exposure and labeled wherever it applies. A candidate is state-dependent if it produces at least two distinct canonical probability vectors on at least two distinct tuning states with the same canonical answer set and displayed order. If the selected local representative is not state-dependent, it is labeled template-derived and the state-dependent candidate chosen by the same rule is reported alongside; test outcomes enter no selection.

Paired contrasts. Each contrast uses the common valid outcomes of the two implementations and reports their count and fraction of all attempts, together with each implementation's all-attempt valid and correct coverage. If either implementation assigns zero probability to a realized event, the NLL difference is undefined and the Brier difference serves as the comparison. Contrasts are formed only within one partition and stratum. On the reduced highD, AIS and weather cohorts, all implementations are restricted to the same outcome and option identifiers, and temperatures and selective thresholds are refitted from each implementation's raw probabilities on the same nested 256-outcome calibration set; this equalizes calibration data only, not training or selection budgets.

Calibration and selective metrics. Calibration error is the sample-weighted absolute difference between bin accuracy and mean confidence over 15 equal-width bins, with empty bins shown. Equal-probability top predictions resolve to the lexicographically smallest event identifier. For a target acceptance fraction, the threshold is the largest observed calibration confidence whose all-attempt acceptance reaches the target; tied confidences are accepted together, and targets above the valid-response rate are reported as unattainable. The area under the risk–coverage curve uses the expected error count over random permutations within confidence ties and is conditional on valid responses. When a fine partition refines a coarse one, its mass is summed over each coarse event before comparison with the directly predicted coarse distribution.

Resampling. Intervals are conditional on fitted models, representatives, temperatures and thresholds. For each family and stratum, 1,000 resamples with seed zero share one set of unit or block multiplicities across methods, calibration variants, option variants and metrics (Table 3). All observations of a dependence unit move together, resampled score sums are divided by the resampled common-valid count, and we report the 2.5th and 97.5th percentiles with the numbers of empty, single-unit and undefined draws, without redrawing. Weather draws rescale block multiplicities within each test year to its original support. No interval is pooled across domains; a secondary summary averages paired Brier differences with equal domain weights, names contributing domains and does not rank models on a common scale.

Sparse support. A temporal, entity, location or spatial axis with fewer than ten test units or five calibration units receives point estimates, unit contrasts and conditional intervals, without interpretation as generalization to a population of periods, regions, recordings, identifiers or owners. Class-specific calibration inference requires at least 50 positive and 50 negative outcomes and five groups containing each class. For highD and weather, domain controls and learned heads are additionally scored at natural prevalence on the complete eligible test sets.

D COST MEASUREMENT PROTOCOL

The protocol measures steady warm service by a persistent local decision service; data reconstruction, model loading, fitting and warm-up are excluded. Measurements use one host with an Intel Core Ultra 7 265H processor (16 logical processors, 63.5 GiB memory, no discrete GPU) in float32, without competing jobs. For each family, the timed implementations are the domain controls, the three local readouts and the seed-zero learned head. Each timed call processes one outcome, including payload construction, encoding, inference, parsing and calibration. Two operating points share eight threads: one request with eight intra-operation threads, and two concurrent requests with four threads each. After an untimed warm-up batch of 32 tuning outcomes, a fixed set of 32 common test outcomes is processed three times in randomized method order, giving 96 timed completions per family, implementation and operating point; we report median, 95th percentile and maximum latency, and throughput as valid completed decisions per second with attempted counts. Every measurement must reproduce the saved predictions within fixed absolute tolerances (10^{-10} for controls, 10^{-6} for language models and learned heads), or the run is rejected.

Energy is gross CPU-package energy per completed request, read from the operating system's cumulative RAPL counter. Each implementation and operating point receives one contiguous window of at least 30 seconds and ten observed counter updates, bracketed by equal-duration idle windows that give a secondary net estimate retained with its sign. Unreadable, reset or non-advancing counters, nonpositive energy or insufficient duration mark a cell unmeasured, and no energy is imputed from power ratings, time or token counts. Two implementations are not ranked by energy when their difference is below the larger observed idle drift. Hosted latency is a separate regime summarized from client-observed request logs, including network transport, rate-limit responses and retries; it measures neither server latency nor server energy. Quality–cost results are reported as points on the same paired workload, and no matched-quality speedup is forced when quality or coverage differs.

E Reproducibility details

The local backbone is Qwen/Qwen3-0.6B at revision `c1899de289a04d12100db370d81485cdf75e47ca`, with 596,049,920 parameters, run with one inter-operation thread and a recorded number of intra-operation threads. Sample selection uses seed zero; the learned head uses seeds zero, one and two; paired resampling uses 1,000 draws with seed zero. The hosted request ledger retains every physical attempt, its status, returned probabilities or failure, and client-observed latency, with provider identifiers and usage where returned; credentials are excluded. Derived partitions record the adapter version and the task, split and feature definitions used to build them. The sources are Coinbase BTC-USD full-channel order-book captures, IEX captures, Bitcoin mempool collector logs, highD (Krajewski et al., 2018), Daymet V4 (Thornton et al., 2021), AIS reports and τ^2-bench traces replayed under a pinned runtime (Barres et al., 2025). Redistribution follows each source's terms; raw highD trajectories are not redistributed and were not sent to the hosted service.

F Additional tables

Family	Tuning	Calibration	Test	Held-out part.	Permutation	Paraphrase	Total
LOB	1,024	1,189	1,024	1,024	256	256	4,773
IEX	270	270	540	540	256	256	2,132
Mempool	549	85	1,024	1,024	256	256	3,194
highD	128	256	256+256	—	32+32	32+32	1,024
Weather	128	256	256+256	256+256	32+32	32+32	1,536
AIS	128	256	256+256	256+256	32+32	32+32	1,536
Agent	224	272	288	—	256	256	1,296

Table 6: Local attempts per readout by split and display pool. "Test" counts base-option test outcomes; entries $a+b$ give the primary and additional test strata (highD location, weather spatial, AIS identifier); held-out partitions, permutations and paraphrases reuse test outcomes. Every attempt of every readout is valid; free and constrained outputs equal the format example in every cell, one-pass outputs in none.