TimeClaw: A Time-Series AI Agent with Exploratory Execution Learning

Hangchen Liu[1] · Dongyuan Li[1]* · Jiewen Deng[2] · Weiwei Ye[1] · Yoshihide Sekimoto[1] · Renhe Jiang[1]

[1] The University of Tokyo [2] Southern University of Science and Technology * Corresponding author

Paper

Learn reusable tool-use strategies from measured execution quality.

1. Learn from execution quality

Research question: how can an agent learn from tool-use paths that all produce valid answers, yet differ in prediction error? Compare executed paths with task-specific training metrics.

$$q(\pi;x) = -\mathcal{L}_\tau(\hat{y}_\pi, y^\star)$$

π: executed path; ŷπ: prediction; y*: training target. Lτ: task loss; Πvalid: paths satisfying the output contract.

$$\pi^\star(x) = \arg\max_{\pi \in \Pi_{\text{valid}}(x)} q(\pi;x)$$

Execute ≥2 distinct valid paths. When experience exists, compare a prior-guided path with an alternative. Retain weaker/failure evidence; use training-only scores.

2. Keep exploration diverse

Tool-prior collapse: familiar tools dominate exploration before alternatives are compared.

$$p_{\text{keep}}(i \mid s) = \left(\frac{1+n_{\min}(s)}{1+n_i(s)}\right)^\alpha, \quad \alpha > 0$$

s: task scope; n_i: historical tool use; n_min: least-used non-protected competitor; α: suppression strength. Protected tools always remain visible.

More-used tools get lower keep probability. First-100 coverage: forecast 22.7% → 72.7%; MACD 15.2% → 65.2%.

3. Distill reusable experience

A reusable rule preserves more than a preferred tool:

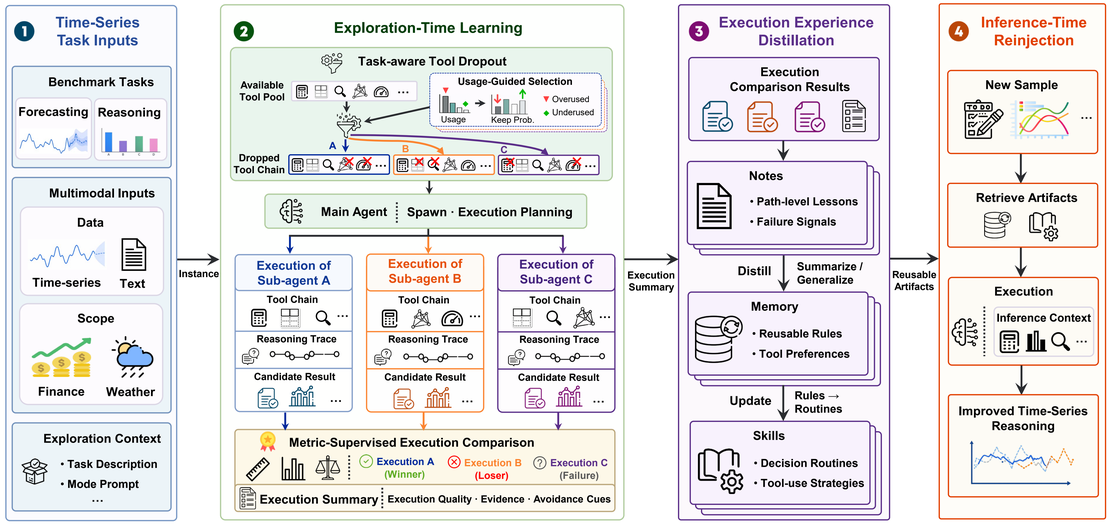
When	Task + applicability conditions
What	Preferred / avoided tool paths
Why	Rationale + supporting episodes
Caution	Confidence + injection status

Notes retain episode evidence. Memory holds rules; tool notes and skills supply usage guidance and procedures.

At inference: retrieve applicable, eligible experience; keep the base model and experience read-only.

4. Explore → Compare → Distill → Reinject

Frozen base agent · training-only comparison · read-only inference



Worked example A · exploration (Appendix G.1, Fig. 7)

Input: 14 days of hourly temperature + boundary storm / funnel-cloud text. Task: future-trend classification.

Path A · text-aware classification

Series + storm context → classifier

decreasing · correct

Path B · forecast → classification

Chronos2 → classifier

increasing · incorrect

Path B: last-day mean 26.631 → forecast-day mean 29.750 (Δ = +3.119). Training label: decreasing.

Recorded lesson: near boundary storms, compare a text-aware route with a forecast route; respect the explicit day-mean rule.

5. Performance: key findings

17 tasks · 30 metrics · best reported on 27/30

Selected task / metric	GPT-5.4	TimeClaw
Weather forecast 7D / MSE ↓	14.197	**10.719**
Finance trend 7D / 5-way acc. ↑	56.9%	**69.6%**

27/30: all listed baselines; benchmark filtering applies. Gaps: finance correlation-long and MACD-30D.

6. Evidence and reuse

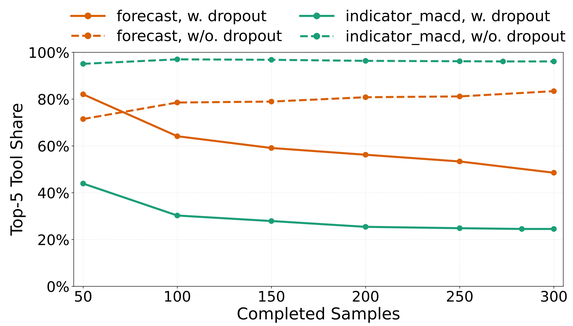
Does distilled experience matter? (Table 7)

Same agent and tools; remove experience injection.

Task / metric	w/o Exp.	Full
Weather FC-S / MSE	13.058	**10.719**
Finance corr-S / acc.-3	56.4%	**58.0%**
Weather QA-L / acc.	59.6%	**66.4%**
Finance trend-S / acc.-5	66.3%	**69.6%**

FC: forecast; S/L: short/long; QA: multiple-choice. MSE: mean squared error; acc.: accuracy.

Does dropout preserve exploration?



Top-5 share: fraction of calls using the five most-used tools. Lower = less concentration (Fig. 4).

Worked example B · inference

Appendix G.2, Fig. 12 · different weather sample

Memory: compare daily means (±0.5); prioritize the last day and boundary storm context.

segment → window_stats

→ llm_reasoning_classification

336 points → 14 daily windows; last-day mean 16.435. Output: increasing = ground truth.

Illustrative traces; not a matched transfer experiment.