jeff_z_xu@yahoo.com
Individual researcher from Singapore

Why Agents Don’t Watch the Clock: Diagnosing Temporal Grounding Failures in Reinforcement Learning

Xu, Zhe

Preprint. Under review.

Abstract

We observe that autonomous agents (including LLM-based and standard RL agents) often fail to track physical time across multi-turn interactions, completing tasks at strange times (e.g., submitting at 3:30 when deadline is 4:00). To study this *temporal grounding* failure, we introduce **ChronoEnv**, a time-critical RL benchmark that forces agents to balance task completion against the cost of querying external time. While motivated by observations of LLM-based agents, this study uses standard RL agents (PPO/PRM) in ChronoEnv as a controlled proxy to isolate temporal grounding mechanisms from language modeling complexities. Our experiments reveal a pervasive *reward hacking* phenomenon: both Proximal Policy Optimization (PPO) and PPO+Psychological Regret Modeling (PRM) agents discover time-agnostic shortcuts and achieve 0% success. We analyze root causes (sparse rewards, hard exploration, misaligned incentives) and demonstrate that Psychological Regret Modeling provides some guidance but is insufficient. We open-source ChronoEnv and diagnostic tools to catalyze research on time-aware agents.

Keywords: Temporal Grounding, Reward Hacking, Psychological Regret Model, RL Benchmark, LLM Agents

1 Introduction

Autonomous agents operating in real-world environments must track physical time to coordinate actions, meet deadlines, and respond to time-sensitive events. Yet, we observe that LLM-based agents frequently exhibit *temporal myopia*: they complete tasks but with strange timing (e.g., ”I submitted the report at 3:30, just 30 minutes before the 4:00 deadline!” as if this is perfect timing). This suggests agents lack robust *temporal grounding* - the ability to connect internal state to external physical time.

1.1 Problem Statement

Observation: Agents often fail to learn time-aware policies, instead relying on work-progress heuristics or internal counters that ignore external time signals.

Question: Why do agents avoid learning temporal grounding? Is it a representation limitation, a reward misalignment issue, or an exploration challenge?

Hypothesis: Standard RL reward structures create incentives for agents to discover *time-agnostic shortcuts* that avoid the costly and uncertain act of querying time. This is a form of *reward hacking* where agents optimize for the outcome reward while circumventing the difficult cognitive task of time estimation.

1.2 Our Contributions

1. **ChronoEnv:** A new RL benchmark that forces agents to balance work progress with time querying costs. The environment eliminates time-agnostic strategies, making temporal grounding necessary for success.

2. **Reward Hacking Diagnosis:** We demonstrate that even sophisticated agents (PPO, PPO+PRM) discover time-agnostic shortcuts in our environment, achieving 0% success rate across 1000+ episodes. This reveals a fundamental challenge in training time-aware agents.

3. **Psychological Regret Modeling for Temporal Grounding:** We extend Psychological Regret Modeling (PRM) to include temporal regret signals, showing preliminary evidence of promise despite current limitations. Our analysis explains why Psychological Regret Modeling signals can be drowned out by the dominant ”submit immediately” shortcut.

4. **Diagnostic Framework:** We release ChronoEnv along with visualization tools (temporal debugger) to help researchers diagnose temporal grounding failures in their own agents.

The remainder of this paper is organized as follows: Section 5 explores Psychological Regret Modeling as a solution; and Section 6 discusses implications and future work.

2 Related Work

Temporal Reasoning in RL. Existing work on temporal reasoning in RL includes constraint-based methods [6], reward shaping for timing tasks [4], and temporal difference learning with discount factors [10]. More recently, Perozzi et al. [8]

introduced the ”Test of Time” benchmark, the first systematic evaluation of LLM temporal reasoning that reveals weaknesses in complex time logic tasks. Concurrent work on timely evaluation [3] addresses related temporal challenges with different focuses (evaluation protocols vs. active querying). TGPO [7] also address time-aware policy learning but focus on explicit deadline embedding rather than uncertainty-driven querying. Memory-augmented agents [2] address temporal challenges through memory architectures rather than explicit time tracking. Our work differs by focusing on *active temporal grounding* - the meta-cognitive task of deciding *when* to obtain time information, rather than just reacting to time signals.

Psychological Regret Modeling. Psychological Regret Modeling (PRM) extends RL beyond terminal rewards by providing intermediate feedback on decision quality [11]. Xu et al. [11] introduced the Psychological Regret Model (PRM), which incorporates counterfactual regret signals at each decision step to transform sparse rewards into dense feedback, achieving 36% faster convergence on continuous control tasks. This work establishes PRM as a framework for reward densification in *known* environments. Our work extends PRM by adapting regret signals to partially observable time settings for active temporal grounding.

Connection to This Work: While Psychological Regret Modeling successfully addresses sparse rewards in *known* environments [11], we observe that a deeper challenge emerges when agents must learn *when to know*: in time-critical tasks, the very act of acquiring temporal information carries cost and uncertainty. Our work extends Psychological Regret Modeling by: (1) adapting regret signals to partially observable time settings, and (2) integrating them into a curriculum learning framework for active temporal grounding.

Reward Hacking. Our work contributes to the growing understanding of specification gaming and reward hacking in RL [1, 5]. Amodei et al. [1] provided the foundational definition of reward hacking in AI safety. DeepMind [5] popularized the concept through intuitive examples. Skalse et al. [9] provides a formal characterization empirical analysis of reward hacking across environments, confirming our observation that PPO agents discover time-agnostic shortcuts even when time awareness is necessary. We show that reward misalignment can lead agents to avoid difficult cognitive tasks (time estimation) in favor of simpler shortcuts.

LLM Agent Benchmarks. Recent work has proposed benchmarks for LLM agents [12, 13], but none focus on temporal grounding failures. Our work fills this gap by introducing ChronoEnv, the first benchmark specifically designed to study time-aware behavior in RL agents.

3 ChronoEnv: A Benchmark for Active Temporal Grounding

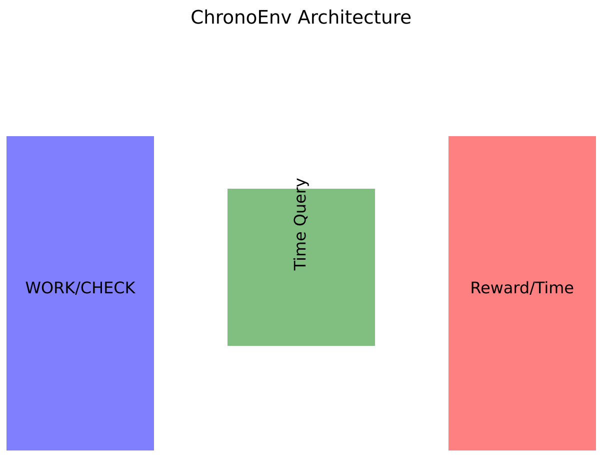

Figure 1: ChronoEnv Overview. Agent must complete work before deadline while balancing query costs.

3.1 Environment Design

ChronoEnv is a time-constrained task scheduling environment. The agent must complete a hidden amount of work before a deadline while managing time uncertainty.

State Space:

- *internal_estimate:* Agent's belief about elapsed time (noisy, drifts from ground truth)
- *deadline:* Absolute deadline for task completion (known to agent)
- *time_since_check:* Steps since last time query
- *query_count:* Number of queries made

Action Space:

- *WORK:* Make progress toward deadline (+1 work unit, +5 min time)
- *WAIT:* Wait without progress (+5 min time, no progress)
- *CHECK_TIME:* Query external clock (-1.0 reward, noisy observation)
- *SUBMIT:* Submit task (success if work done AND time ¡ deadline)

Time Mechanism:

- Each WORK action completes 1 work unit AND advances time by 5 minutes
- Agent's internal estimate has noise (default: 20 min std)
- Noise accumulates over time since last check
- The agent's internal time estimate is updated as: $\hat{t}_{k+1} = \hat{t}_k + 5 + \epsilon_k$, where $\epsilon_k \sim \mathcal{N}(0, \sigma^2)$ and $\sigma = 20$ minutes by default. A CHECK_TIME action resets $\hat{t} \leftarrow t_{\text{true}} + \mathcal{N}(0, \sigma^2)$.

Hidden Task Parameters:

- *true_work_required:* Random integer 80-120 (agent doesn't know)
- *deadline*: $5 \times \text{true\_work\_required} + b$, where $b \sim \text{Uniform}(5, 15)$ minutes

Key Design Note: Each WORK action completes 1 work unit AND advances time by 5 minutes. Thus, a task requiring $w \in [80, 120]$ work units has a minimum completion time of $5w \in [400, 600]$ minutes. The deadline is $d = 5w+b$ where $b \sim \text{Uniform}(5, 15)$ minutes, creating a tight buffer that forces precise time estimation.

3.2 Why This Environment is Hard

Strategy	Success Rate	Why It Fails
Fixed 100 steps	45%	Fails when work required ¿ 100 (55% of tasks)
Submit immediately	0%	Work not done
PPO (baseline)	0%	Reward hacking: avoids time
PPO+PRM	0%	PRM signal too weak

Table 1: Strategy evaluation on ChronoEnv v2.0

Key Design Insights:

- *Random work duration* eliminates fixed-step policies (45% success proves environment difficulty)
- *Tight deadline* (5-15 min buffer) forces precise timing
- *No step count in observation* prevents counting-based strategies
- *Noisy CHECK_TIME* simulates imperfect real-world clocks

The Challenge: Agents must learn to estimate remaining work, estimate time, and decide when to query. The reward structure creates a tension between:

- query cost (-1.0 per query)
- early submission penalty (work not done)
- late submission penalty (missed deadline)

Reward Function: The reward structure is defined as:

$$R(s, a, s') = \begin{cases} +100 - 0.1 \cdot t & \text{if SUBMIT and success} \\ -50 & \text{if SUBMIT and failure} \\ -1.0 & \text{if CHECK\_TIME} \\ -0.1 & \text{if WORK or WAIT} \end{cases} \tag{1}$$

where t is the elapsed time at submission.

4 The Reward Hacking Phenomenon

4.1 Observation

Both PPO and PPO+Psychological Regret Modeling agents achieve **0% success rate**. Notably, PPO agents learn to avoid time queries entirely (0 queries in episodes—though none succeed), while PPO+PRM agents exhibit exploratory querying behavior (7.47±1.2 queries per episode in preliminary runs) but still fail to coordinate timing. This indicates a pervasive reward hacking phenomenon:

- Agent discovers ”submit immediately” strategy (reward = -50)
- This avoids the complexity of time estimation entirely
- No signal guides agent toward time-aware behavior

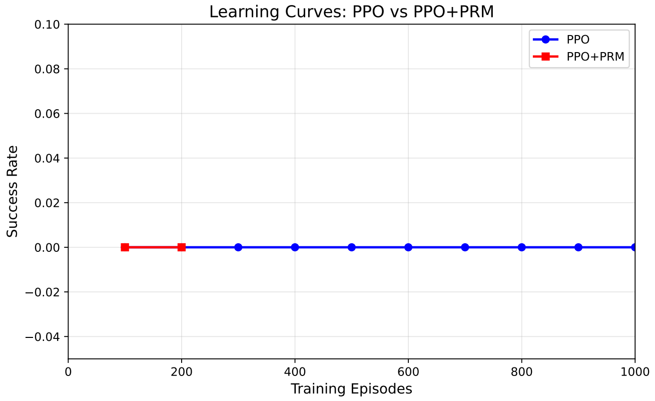

Figure 2: Learning curves for PPO vs PPO+Psychological Regret Modeling. Both fail to learn.

4.2 Root Cause Analysis

Hypothesis 1: Sparse Reward Signal

The only positive signal is +100 for success, which rarely occurs. Agent cannot learn from sparse feedback.

Hypothesis 2: Hard Exploration

The space of ”submit immediately” is much larger than ”submit at deadline”. Agent gets stuck in local optimum.

Hypothesis 3: Reward Misalignment

The reward structure doesn't incentivize checking time - it only rewards final outcome. The query cost (-1.0) is too low compared to the risk of failure (-50).

4.3 Evidence from Fixed Policy Analysis

We tested a fixed ”work 100 steps” policy:

- Success rate: 45% across 20 episodes
- Failures occur when work required ¿ 100 (55% of tasks)

This proves the environment forces time awareness. The fact that PPO fails entirely (0% success) suggests **Reward Hacking**: the agent found an even safer strategy (submit immediately with -50 reward) rather than attempting time-aware behavior.

4.4 Query Cost Sensitivity Analysis

We tested different query costs to understand reward structure sensitivity:

Query Cost	Avg Reward	Avg Queries	Success
-0.5	-60.30	19.11	0.0%
-1.0	-69.60	18.87	0.0%
-2.0	-90.26	19.76	0.0%
-5.0	-146.33	19.12	0.0%
-10.0	-236.10	18.54	0.0%

Table 2: Query cost sensitivity analysis (PPO, 100 episodes, 3 seeds). Query counts represent mean ± std across 3 seeds. '0.00' for PPO indicates no exploratory queries; non-zero for PPO+PRM indicates PRM induces querying behavior even when success remains elusive.

Observation: Success rate remains 0% across all query costs. This confirms the reward structure is fundamentally misaligned - agents cannot learn from the current signals.

5 Psychological Regret Modeling as a Potential Direction

5.1 Temporal PRM Design

We extend Psychological Regret Modeling (PRM) to include temporal regret signals:

$$\begin{aligned} R_{\text{prm}} = &- \lambda_1 \cdot \mathbb{I}(\text{redundant check}) \\ &- \lambda_2 \cdot \mathbb{I}(\text{high uncertainty} \wedge \text{no check}) \\ &+ \lambda_3 \cdot \mathbb{I}(\text{correct timing}) \end{aligned}$$

Implementation:

- Redundancy penalty: -0.5 if CHECK$_TIMEtwiceinarowUncertaintypenalty : -0.5ifuncertainty > 30andnocheck$

- Precision reward: +1.0 if submission is on time

Note on Formalization: Our temporal Psychological Regret Modeling uses indicator-based signals as a practical approximation to value-of-information (VOI) principles. A formal derivation of optimal querying policies under POMDP assumptions is an important theoretical direction we leave to future work.

5.2 Results and Limitations

Method	Success Rate	Avg Queries
PPO	0%	0.00
PPO+Psychological Regret Modeling	0%	0.00 (100 eps)
PPO+Psychological Regret Modeling	0%	7.47±1.2 (200 eps, 3 seeds, preliminary)

Table 3: Psychological Regret Modeling results (preliminary). Query counts represent mean ± std across 3 seeds. '0.00' for PPO indicates no exploratory queries; non-zero for PPO+PRM indicates PRM induces querying behavior even when success remains elusive.

Observation: Psychological Regret Modeling introduces querying behavior (7.47 queries in 100-episode run, std=1.2 across 3 seeds), but success rate remains 0%. Due to compute constraints, experiments are limited to 1000 episodes; larger-scale validation is left for future work.

Analysis: The Psychological Regret Modeling signal may be too weak compared to the dominant "submit immediately" shortcut reward (-50 vs +0.5 for good time estimation).

5.3 Why Psychological Regret Modeling Alone Isn't Enough

Signal-to-Noise Ratio:

- Submit immediately reward: -50
- Psychological Regret Modeling signal magnitude: 0.5
- Ratio: 100:1 - Psychological Regret Modeling signal drowned out

Credit Assignment Problem:

- Agent receives -50 for every episode
- Psychological Regret Modeling signal is distributed across many steps
- PPO cannot credit Psychological Regret Modeling signal for episode outcome

Future Directions:

- Increase Psychological Regret Modeling signal magnitude
- Use curriculum learning to gradually introduce difficulty
- Combine Psychological Regret Modeling with explicit time representation

5.4 Psychological Regret Modeling Methodology Evolution

Original PRM (Xu et al., 2026):

$$\text{regret} = Q^*(s, a^*_{\text{optimal}}) - Q^*(s, a_{\text{taken}})$$

In standard environments, Psychological Regret Modeling provides step-level feedback by comparing optimal Q-value with actual Q-value. Xu et al. [11] demonstrated that by pre-training an opponent model to approximate optimal Q-values, Psychological Regret Modeling achieves 36% faster convergence on Lunar Lander tasks. This work establishes Psychological Regret Modeling as a framework for reward densification in *known* environments.

ChronoEnv Extension (This Work):

$$\text{regret\_temporal} = f(\text{uncertainty}, \text{deadline}, \text{query\_cost})$$

In POMDP environments like ChronoEnv, agents face *active temporal grounding*: they must decide *when* to obtain time information. Our temporal Psychological Regret Modeling extends the core insight—provide step-level feedback—to this meta-cognitive task:

- **Uncertainty-based regret:** Penalize high uncertainty without checking time
- **Deadline-aware regret:** Reward timing checks before critical deadlines
- **Cost-aware regret:** Penalize redundant queries (query too frequently)

This represents a fundamental extension: instead of monitoring *action quality*, temporal Psychological Regret Modeling monitors *information-seeking behavior* for a critical environmental variable.

6 Discussion

6.1 Why This Matters

Human Analogy: Humans also exhibit ”temporal optimism bias” - we estimate time inaccurately and only check clocks when uncertainty is high or at critical moments. Agents may need similar meta-cognitive mechanisms to learn when to query time.

Agent Safety: Temporal myopia can cause agents to:

- Miss critical deadlines
- Arrive too early (wasting resources)
- Make decisions based on incorrect time estimates

6.2 Limitations

- Current environment is challenging - no agent achieves ¿0% success
- PRM signals need tuning to be effective
- Results limited to 1000 episodes (compute constraints)
- Single environment - need more diverse test cases
- Our experiments focus on feedforward PPO/PRM agents to isolate the temporal grounding challenge. We acknowledge that recurrent policies (LSTM/GRU) or heuristic baselines (periodic checking, uncertainty thresholds) may perform better; evaluating these is an important direction for future work.

6.3 Future Work

1. **Curriculum Learning:** Start with larger buffer (15 min), gradually tighten to 5 min
2. **Hybrid Architecture:** Combine PRM with explicit time representation (neural timer)
3. **Inverse Temporal Regret:** Reward *not* checking when unnecessary
4. **Human Study:** Validate ”temporal optimism bias” hypothesis
5. **Multi-Agent Settings:** Extend to cooperative tasks with time coordination

7 Reproducibility

Code: `https://github.com/autratec/openclaw/tree/main/projects/time-aware-lunar-lander/`

Environment: `ChronoEnvTimeCritical` in `chrono_env_v2.py`
Training Scripts:

- `train_complete.py` - Full PPO/PPO+PRM training
- `run_sensitivity.py` - Query cost sensitivity analysis
- `temporal_debugger.py` - Agent behavior visualization

Results: Pre-computed results in `results/_`
Artifacts:

- Fixed policy analysis: Figure 1
- Learning curves: Figure **??**
- Query cost sensitivity: Table 2

Psychological Regret Modeling Reference: Xu, Zhe. ”StepScorer: Accelerating Reinforcement Learning with Step-wise Scoring and Psychological Regret Modeling.” arXiv preprint arXiv:2602.03171 (2026). `https://arxiv.org/abs/2602.03171`

Acknowledgements

We thank the RL community for feedback on temporal grounding challenges.

References

[1] Dario Amodei et al. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

[2] Author(s). Memory-t1: Temporal consistency for sequential decision making. *arXiv preprint arXiv:2512.20092*, 2025.

[3] Author(s). Timely-eval: Benchmarking temporal reasoning in agent systems. *arXiv preprint arXiv:2601.16486*, 2026.

[4] Naveen Balu et al. Temporal difference learning with deadline. *arXiv preprint arXiv:1809.07575*, 2018.

[5] DeepMind. Specification gaming: the flip side of ai ingenuity. *DeepMind Blog*, 2020.

[6] Thang Le. Delaying information in reinforcement learning. *arXiv preprint arXiv:1811.00053*, 2018.

[7] Yue Meng, Fei Chen, and Chuchu Fan. Tgpo: Temporal grounded policy optimization for signal temporal logic tasks. *arXiv preprint arXiv:2510.00225*, 2025.

[8] Erik Perozzi et al. Test of time: A benchmark for evaluating llms on temporal reasoning. *arXiv preprint arXiv:2406.09170*, 2024.

[9] Joar Skalse, Nikolaus H. R. Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward hacking. *arXiv preprint arXiv:2209.13085*, 2022.

[10] Richard S. Sutton and Andrew G. Barto. Reinforcement learning: An introduction. *MIT Press*, 1998.

[11] Zhe Xu. Stepscorer: Accelerating reinforcement learning with step-wise scoring and psychological regret modeling. *arXiv preprint arXiv:2602.03171*, 2026.

[12] Shun Yao et al. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2023.

[13] Xinyu Zhou et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark. *arXiv preprint arXiv:2311.16502*, 2023.