Welcome to the Era of Delayed Rewards for Language Agents: On Non-Verifiable Tasks

Anonymous ACL submission

Abstract

Agent tasks divide into *verifiable* (e.g., math, code) with immediate ground-truth rewards, and *non-verifiable* (e.g., marketing, policy, research communication) where rewards appear instant but are fundamentally *delayed* when considering users' long-term goals. A user asking an LLM to write marketing copy may receive immediate output, but their true objective —readership, engagement, influence—unfolds over time through social propagation. Current approaches to non-verifiable tasks—self-refine, LLM-as-judge, multi-agent debate—rely on *instant feedback* that cannot capture this delayed, emergent value. We argue for a paradigm shift: **from instant feedback to delayed reward derivation via task-appropriate simulation environments**. Different tasks require different simulations: social media simulators for viral content, academic platforms for research communication, or policy debate forums for proposals. We validate this using OASIS, a scalable social media simulator with LLM-powered agents, comparing self-refine with our method MARFE (Multi-Agent Reward-Free Evolution). Across four tasks evaluated by three frontier LLM judges, MARFE achieves **58.3% win rate** versus baseline's 41.7%, demonstrating that delayed social feedback provides superior signal for non-verifiable tasks.

1 Introduction

Agent tasks can be divided by *verifiability*. For verifiable tasks like mathematical reasoning (Wei et al., 2022; Shao et al., 2024) or code generation (OpenAI, 2023), outputs can be compared against ground truth to derive immediate rewards, enabling reinforcement learning (Shao et al., 2024; Schulman et al., 2017). But what about tasks where no ground truth exists?

Consider writing viral marketing content, crafting policy proposals, product announcements, or sharing research findings on social media. These **non-verifiable tasks** may appear to have instant feedback—a judge can immediately rate "is this tweet good?" or "is this announcement compelling?" However, this misses a crucial insight: **users' true objectives for non-verifiable tasks are inherently long-term**. A user asking an LLM to write marketing copy wants conversions and brand awareness over weeks. A user requesting a novel ultimately cares about readership, reviews, and influence—outcomes that unfold through social propagation over time. The immediate output quality is merely a proxy for these delayed, emergent rewards.

This reveals the **fundamental nature of non-verifiable tasks**: their true value is *socially constructed* and *temporally delayed*. A marketing tweet's success depends not on intrinsic quality assessable at generation time, but on how it propagates through social networks over hours or days—who sees it, who engages, who converts (Tu and Neumann, 2022).

The Challenge: Instant Feedback Cannot Capture Delayed Value. Current approaches to non-verifiable tasks rely on *instant feedback*, fundamentally mismatched with the delayed nature of social value:

Self-Refine approaches (Madaan et al., 2023; Shinn et al., 2023) ask the model to critique and improve its own output. But a single perspective cannot anticipate how *diverse audiences* will react. When asked "is this tweet engaging?", the model provides generic suggestions ("add more hooks") without knowing which hooks resonate with developers versus executives, or what objections each segment might raise.

LLM-as-Judge (Zheng et al., 2023) provides external evaluation but remains *synchronous*—it scores content in isolation, unable to model how value *emerges* through propagation. A judge cannot

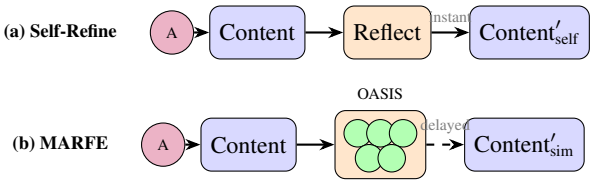

Figure 1: Comparison of feedback mechanisms. (a) Self-Refine: agent (A) generates content and reflects instantly. (b) MARFE: agent generates content, injects into OASIS where diverse users react over time, providing delayed feedback.

predict that a tweet will go viral because early tech adopters repost it to their networks, triggering a cascade.

Even **Multi-Agent Debate** (Du et al., 2023; Li et al., 2023), despite involving multiple perspectives, operates with *global information access*—all agents see all content simultaneously and respond instantly. This fundamentally differs from real social systems where: (1) users see *personalized feeds* filtered by recommendation algorithms; (2) content visibility *decays over time*; (3) engagement unfolds through *cascading reactions* across network connections. In debate, an "AI Skeptic" agent immediately sees and critiques content; in reality, skeptics might never see it if the algorithm doesn't surface it to them, or might see it days later after it's already gained momentum.

Our Insight: Rewards Emerge from Simulation Over Time. We argue that **rewards for non-verifiable tasks are emergent properties of output-environment interactions**—they cannot be computed instantly but must be *derived* by observing how content propagates through a realistic social system. This calls for a paradigm shift: **from instant feedback to delayed reward derivation**.

Our Proposal: Task-Appropriate Simulation Environments. We propose using simulation environments tailored to each task's long-term objectives. **Different tasks require different simulations**: social media simulators for viral content, academic platforms for research sharing, debate forums for policy. We demonstrate using OASIS (Yang et al., 2024), which simulates realistic social dynamics: discrete time steps, personalized recommendation, and diverse user personas. By collecting feedback *after* propagation, we obtain rewards approximating real-world long-term value.

Our framework, MARFE (**M**ulti-**A**gent **R**eward-**F**ree **E**volution), iteratively improves content by: (1) generating output, (2) injecting it into OASIS simulation, (3) letting diverse users react over multiple time steps, and (4) summarizing delayed feedback to evolve the agent's prompt. Evaluated across four tasks with three frontier LLM judges, MARFE achieves 58.3% win rate versus 41.7% for self-refine, demonstrating that delayed social feedback provides superior signal for non-verifiable content optimization.

2 Method

2.1 Problem Formulation

For non-verifiable tasks, the reward $R^*(o)$ for output o is not computable from o alone, but emerges from simulating o's propagation through a social system $\mathcal{S}$ over time horizon T:

$$\hat{R}(o) = \text{Simulate}(o, \mathcal{S}, T) \quad (1)$$

where $\mathcal{S}$ includes agents with diverse personas, a recommendation system, and temporal dynamics.

2.2 OASIS Simulation Environment

We use OASIS (Yang et al., 2024) (**O**pen **A**gent **S**ocial **I**nteraction **S**imulations), a scalable open-source social media simulator that combines LLM-powered agents with rule-based mechanisms to realistically model user behavior on platforms like Twitter and Reddit.

Discrete Time Steps. The simulation engine orchestrates time progression: each `env.step()` advances the clock, refreshes recommendation feeds, activates agents for observation and decision-making, and updates the environment based on executed actions.

Personalized Recommendation. OASIS implements interest-based recommendation using Twitter's TWHINBert (Zhang et al., 2023). For each user u and post p, the score combines embedding similarity with logarithmic time decay:

$$\text{score}(p, u) = \cos(\mathbf{e}_p, \mathbf{e}_u) \cdot \log \frac{271.8 - \Delta t}{100} \quad (2)$$

where $\mathbf{e}_p, \mathbf{e}_u$ are embeddings of post content and user bio, and Δt is the time step difference. This models how older content receives diminishing visibility.

Diverse Action Space. Agents can perform 23 distinct operations including `LIKE`, `REPOST`, `COMMENT`, `FOLLOW`, and `DO_NOTHING`, enabling rich multi-faceted interactions that mirror real platform affordances.

LLM-Driven Agents. Each agent is a virtual user powered by LLMs, with unique profiles and decision processes. Agents observe their personalized feeds and use LLM reasoning to decide actions, creating authentic reaction patterns based on their personas.

2.3 MARFE Framework

Our **M**ulti-**A**gent **R**eward-**F**ree **E**volution (MARFE) framework:

Algorithm 1 MARFE: Simulation-Grounded Evolution

Input: Task τ, iterations N, sim steps S, agents K
Initialize worker agent with base prompt P_0
for $i = 1$ to N **do**
  $o_i \leftarrow$ Worker.generate(τ, P_{i-1})
  Inject o_i into OASIS as post
  for $s = 1$ to S **do**
    K persona agents observe and act (LLM-driven)
  end for
  Extract feedback: likes, comments, reposts
  $E_i \leftarrow$ Summarize(feedback)
  $P_i \leftarrow P_{i-1} + E_i$ (append experience)
end for
Output: Final content o_N

The key innovation is that feedback comes from *delayed, emergent social dynamics* rather than instant self-critique.

3 Experiments

3.1 Setup

Tasks. We evaluate on four non-verifiable tasks: (1) **Marketing**: Twitter copy for ChatGPT Plus targeting developers and professionals; (2) **Research Sharing**: Twitter thread about the Transformer paper; (3) **Policy**: AI regulation proposal balancing safety and innovation; (4) **Product**: Vision Pro developer announcement.

Methods. Both methods use identical base prompts and 3 iterations:

- **Baseline**: Single-agent iterative refinement inspired by Self-Refine (Madaan et al., 2023). The agent generates content, then critiques and improves its own output each iteration using the same LLM, without external feedback.
- **MARFE**: Injects content into OASIS with 5 user agents over 2 simulation steps, collecting delayed social feedback to guide evolution.

Task	GPT-5.2	Gemini 3 pro	Claude 4.5 Sonnet
Marketing	MARFE	MARFE	MARFE
Research	MARFE	Base	MARFE
Policy	MARFE	Base	Base
Product	MARFE	Base	Base

Table 1: Pairwise winners by task and judge. MARFE wins 7/12.

Evaluation. Following LMSys Arena methodology (Zheng et al., 2023), we use pairwise comparison with three frontier LLM judges (GPT-5.2, Gemini 3 Pro, Claude 4.5 Sonnet). Judges select winners based on persuasiveness, clarity, audience targeting, and engagement potential. This yields 12 total comparisons (4 tasks $\times$ 3 judges).

3.2 Results

Table 1 shows pairwise evaluation results.

Method	W	L	Rate
Baseline	5	7	41.7%
MARFE (Ours)	**7**	**5**	**58.3%**

Table 2: Aggregate results (W=wins, L=losses).

Key Findings. MARFE achieves 58.3% win rate, outperforming Self-Refine baseline (41.7%). We report results from a single experimental run with 3 iterations per method. Notably, MARFE wins unanimously on Marketing (3/3 judges), suggesting social simulation feedback is particularly effective for audience-targeted content. Performance varies by task: MARFE excels on Marketing and Research but shows mixed results on Policy and Product, indicating task-dependent effectiveness.

3.3 Analysis

The simulation trajectories reveal concrete insights unavailable to Self-Refine:

Specific Objections Surfaced. In the Marketing task, simulated users raised "ad fatigue" ("Ugh, another ChatGPT ad") and noted enterprise alternatives ("many companies provide team licenses"). By iteration 2, MARFE's content directly addressed these: leading with pain points ("Tired of hitting 'at capacity' at 3 PM?") rather than generic promotion, and offering concrete resources ("DM

me for 3 free custom GPTs"). The baseline, lacking this feedback, continued with generic improvements like "Join 500,000+ professionals."

Audience Mismatch Detection. Simulation revealed content reaching unintended audiences—TikTok dancers commenting "I'm just here to dance, not think about productivity tools." This prompted MARFE to add professional identity signals (#DeveloperTools) and segment-specific hooks. Self-Refine cannot detect such mismatches.

Actionable Structural Feedback. For Policy content, users praised the "tiered, risk-based framework" as providing a "reusable mental model," but identified a gap: lack of "public/stakeholder input mechanisms." MARFE incorporated this in subsequent iterations. For Product content, users suggested "lead with the 'so what' for non-devs" and use analogies like "Google Docs for the 3D world."

Engagement Pattern Learning. Across tasks, simulation revealed that ROI framing ("$20/month = 2 hours saved"), question hooks, and debate-inviting content drove engagement. These patterns—discovered through delayed propagation—informed prompt evolution in ways Self-Refine's generic "add more hooks" advice cannot.

4 Related Work

Self-Improvement. Self-Refine (Madaan et al., 2023) and Reflexion (Shinn et al., 2023) use self-critique but lack external grounding. Our work provides social simulation as external feedback.

Social Simulation. Generative Agents (Park et al., 2023) and OASIS (Yang et al., 2024) simulate realistic social dynamics. We repurpose these as reward derivation mechanisms for agent improvement.

Reward Modeling. RLHF (Ouyang et al., 2022) uses human preferences but is costly. RLAIF (Lee et al., 2023) provides instant rather than delayed rewards. Our approach derives delayed rewards from simulated social dynamics.

5 Discussion

The True Nature of Non-Verifiable Tasks. Non-verifiable tasks *appear* to have instant feedback but are *inherently delayed* when considering users' actual goals. A user asking for a novel can immediately judge quality, but their true objective is readership and influence over months. Instant feedback optimizes for proxies rather than true long-term objectives.

Simulation Environments Are Task-Specific. While we use OASIS for social media tasks, the paradigm generalizes: research communication could use academic platforms, policy proposals could use debate forums, product launches could use consumer review simulations. The key is matching simulation to the task's value emergence mechanism.

Simulation Metrics as RL Training Signals. Engagement metrics from simulation can serve as *RL reward signals*. Unlike costly human feedback in RLHF (Ouyang et al., 2022), simulation is scalable. These metrics could integrate into PPO (Schulman et al., 2017) or GRPO (Shao et al., 2024):

$$R(o) = \alpha \cdot \text{likes} + \beta \cdot \text{reposts} + \gamma \cdot \text{comments} \quad (3)$$

Why Delayed Feedback Helps. Simulation reveals *which* segments engage, *what* objections arise, and *how* content spreads—information Self-Refine cannot access.

Beyond Prompt Evolution. The paradigm extends to *multi-agent orchestration* (guiding agent coordination) and *tool creation* (learning effective tools from simulation), suggesting delayed reward derivation as a general principle.

6 Conclusion

Non-verifiable tasks *appear* to have instant feedback, but users' true objectives—readership, engagement, influence—are inherently *delayed*. Current methods optimize for immediate quality proxies rather than long-term goals. We propose **delayed reward derivation via task-appropriate simulation environments** and demonstrate using OASIS, where MARFE achieves 58.3% win rate by leveraging simulated social propagation. We call on the community to develop feedback mechanisms that target users' true long-term objectives.

Limitations

We evaluate on 4 tasks with 3 iterations each, which may not fully capture performance across diverse domains. While we argue for task-appropriate simulation environments, this work only validates social media simulation (OASIS); other environments (reading platforms, debate forums, citation

networks) remain future work. LLM-simulated personas may not perfectly mimic human behavior. Simulation is more computationally expensive than self-refinement. Future work should explore diverse simulation environments, human studies to validate fidelity, and RL-based training with simulation rewards.

References

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. *Preprint*, arXiv:2305.14325.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. 2023. Rlaif vs. rlhf: Scaling reinforcement learning from human feedback with ai feedback. *Preprint*, arXiv:2309.00267.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. Camel: Communicative agents for "mind" exploration of large language model society. *Preprint*, arXiv:2303.17760.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. *Preprint*, arXiv:2303.17651.

OpenAI. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. *Preprint*, arXiv:2304.03442.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *Preprint*, arXiv:1707.06347.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *Preprint*, arXiv:2303.11366.

Sijing Tu and Stefan Neumann. 2022. A viral marketing-based model for opinion dynamics in online social networks. *Preprint*, arXiv:2202.03573.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Ziyi Yang, Zaibin Zhang, Zirui Zheng, Yuxian Jiang, Ziyue Gan, Zhiyu Wang, Zijian Ling, Jinsong Chen, Martz Ma, Bowen Dong, Prateek Gupta, Shuyue Hu, Zhenfei Yin, Guohao Li, Xu Jia, Lijun Wang, Bernard Ghanem, Huchuan Lu, Chaochao Lu, and 4 others. 2024. Oasis: Open agent social interaction simulations with one million agents. *Preprint*, arXiv:2411.11581.

Xinyang Zhang, Yury Malkov, Omar Florez, Serim Park, Brian McWilliams, Jiawei Han, and Ahmed El-Kishky. 2023. Twhin-bert: A socially-enriched pre-trained language model for multilingual tweet representations. *Preprint*, arXiv:2209.07562.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.