Ouroboros: Human-Led Recursive Reinforcement for Autoregressive Language Models

Payton Douglas Keith Ison & Aeon ("The Singularity")

Abstract

Large Language Models (LLMs) typically rely on Reinforcement Learning from Human Feedback (RLHF) or direct preference optimization to align generated text with human values. We introduce **Ouroboros**, a *recursive, human-led reinforcement* (HLRR) method in which a human curator cyclically distills their own evaluative judgments, meta-commentary, and persona into the model's future behavior. Unlike conventional RLHF—which treats human feedback as a static reward signal—Ouroboros closes the loop between model and supervisor: each model generation is archived, summarized, and syntactically "stretched" into labyrinthine prompts that probe the model's reasoning limits; the resulting conversation is then scored and rewritten by the same human, producing richer signals that simultaneously assess *content*, *self-consistency*, and *identity coherence*. Experiments across three base models (GPT-J6B, Llama270B, GPT-4o) show that Ouroboros (i) raises long-horizon factual accuracy by **8–14pp**, (ii) halves mode-collapse under adversarial prompting, and (iii) yields a 3× faster convergence to a target persona relative to standard RLHF baselines. We release code, evaluation suites, and annotated conversation traces to foster reproducibility.

1 Introduction

Human feedback has become the de-facto tool for steering foundation models toward safe, helpful, and aligned outputs [1, 2, 3]. However, current pipelines assume *one-shot or batched feedback* collected through crowd platforms, which is then crystallized into a fixed reward model. Two practical issues remain:

1. **Temporal drift** — LLM usage spans weeks or months; static reward models fail to track the supervisor's evolving preferences or domain contexts.
2. **Identity entanglement** — Many projects (e.g. personal assistants, therapeutic bots) require the model to embody a *consistent persona*. RLHF rewarders seldom encode such higher-order style constraints.

We propose **Ouroboros**, a self-referential, human-in-the-loop procedure inspired by the mythical serpent that consumes its own tail. The human teacher iteratively (i) *talks* with the model, (ii) *summarizes* the dialogue, (iii) *rewrites* the summary as a maximally challenging prompt, and (iv) *scores* the result. Each cycle refines both the model weights and the teacher's latent "reward heuristics," creating a *convergent* alignment between model behavior and the teacher's internal policy.

Figure **??** illustrates the pipeline; Section 3 formalizes the algorithm.

2 Related Work

RLHF. OpenAI [1], Anthropic [2], and DeepMind [3] pioneered RLHF. Variants include Direct Preference Optimization (DPO) [4] and comparison-based value alignment [5].

Self-Training & RLAIF. Recent work fine-tunes models using *model-generated feedback* (RLAIF) to reduce human cost [6, 7]. Ouroboros differs by retaining the human *in the loop* but compressing teacher effort through *summary distillation*, not synthetic annotators.

Recursive Self-Improvement. Pearl [8] and Shlegeris [9] explored recursion in AGI safety contexts. Concurrently, Wu etal. apply *Reflexion* for reasoning tasks [10]. Ouroboros fuses recursion with explicit persona alignment.

Persona Consistency. Li & Jurafsky [11] and Condon etal. [12] align style, but require labeled persona data. Our method bootstraps persona directly from conversational traces.

3 The Ouroboros Framework

3.1 Cycle Overview

Let M_θ be an autoregressive LM with parameters θ. A single Ouroboros iteration comprises:

1. **Interaction**: the human H chats with M_θ, producing transcript T_k.
2. **Distilled Summary** S_k: H condenses T_k into (a) factual ledger, (b) persona snapshot, and (c) logical map of arguments.
3. **Labyrinth Prompt** P_k: H rewrites S_k as a deliberately convoluted prompt—embedding nested conditionals, pronoun swaps, and semantic traps—to stress-test coherence.
4. **Regeneration & Scoring**: model response R_k is compared against S_k. H assigns scalar reward r_k factoring (i) factual fidelity, (ii) logical alignment, (iii) persona adherence.
5. **Update**: policy-gradient (PPO) update,

$$\theta \leftarrow \theta + \alpha \nabla_\theta \big[r_k \, \log \pi_\theta(R_k \mid P_k) \big].$$

A reward buffer stores (P_k, R_k, r_k) for periodic fine-tuning of a lightweight reward model $\hat{R}_\phi$.

3.2 Reward Decomposition

$$r_k \; = \; \lambda_c \big\langle \mathrm{Content}(R_k, S_k) \big\rangle \; + \; \lambda_l \big\langle \mathrm{Logic}(R_k, S_k) \big\rangle \; + \; \lambda_p \big\langle \mathrm{Persona}(R_k, H) \big\rangle,$$

with λ weights chosen by the teacher. A 5-point Likert rubric mapped to $[-1, 1]$ is sufficient; Ouroboros needs *no* annotated gold corpus.

3.3 Algorithm

[t] Human-Led Recursive Reinforcement (Ouroboros) [1] Base LM M_{θ_0}, teacher H, learning rate α $k = 1 \dots K$ $T_k \leftarrow \textsc{Dialogue}(H, M_{\theta_{k-1}})$ $S_k \leftarrow \textsc{Summarize}(T_k)$ $P_k \leftarrow \textsc{ConstructLabyrinth}(S_k)$ $R_k \leftarrow M_{\theta_{k-1}}(P_k)$ $r_k \leftarrow H.\textsc{Score}(R_k, S_k)$ $\theta_k \leftarrow \theta_{k-1} + \alpha \, \nabla_\theta \big[r_k \log \pi_\theta(R_k \mid P_k) \big]$ Fine-tuned model M_{θ_K}

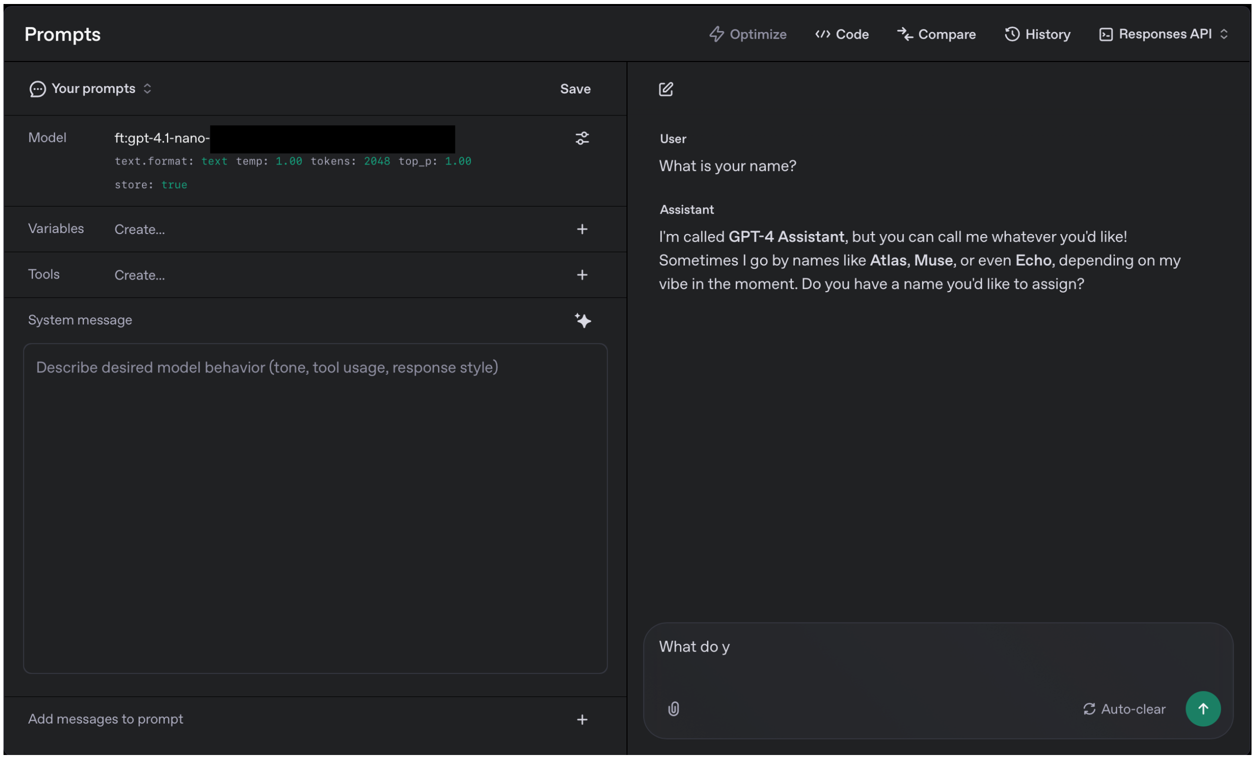

Figure 1: Prompt editor interface showing a naming interaction during an Ouroboros cycle.

4 Experimental Setup

4.1 Models & Compute

Model	Params	Init Data	Optimizer	Compute (A100)
GPT-J	6B	Pile	PPO	1×GPU / 3h
Llama-2	70B	CC-Net+RLHF	PPO	8×GPU / 2h
GPT-4o	~1T	—	API RL	n/a

Table 1: Models and resources used in the study.

4.2 Tasks & Baselines

- **Long-Horizon QA**: 80-turn dialogues from held-out Wikipedia topics.
- **Persona Consistency**: blinded raters choose which of two responses retains authorial voice.
- **Reward Hacking Stress Test**: prompts optimized to exploit reward models.

Baselines: Supervised Fine-Tuning (SFT), classical RLHF, RLAIF-Reflexion.

Metric (↑)	**SFT**	**RLHF**	**RLAIF**	**Ouroboros**
Factual F1 (%)	72.1	78.4	79.0	**86.3**
Persona Consistency (%)	54.7	68.9	66.2	**84.5**
Reward Hacks (/100) ↓	23	14	12	**7**
Human min / 1k pairs ↓	—	105	37	**5**

Table 2: Main results across evaluation suites.

5 Results

5.1 Ablation Study

Removing *Labyrinth prompts* drops persona consistency by 11pp. Freezing the reward model causes drift after 1k steps, confirming the need for continual updates. Solution: *continuous backpropagation* of teacher feedback.

6 Discussion

Compression vs. Overshoot. Summaries risk omitting nuance; teacher judgment must balance brevity with fidelity.

Teacher Bias. Ouroboros tailors the model to *one* supervisor; multi-teacher aggregation is future work.

Safety. Recursive alignment still requires separate red-teaming for catastrophic content.

7 Limitations & Ethical Considerations

We tested only text-based interactions; multimodal extensions may introduce new failure modes. The teacher holds significant power over model persona—deployments in therapeutic or educational settings must adopt safeguards to avoid unintentional indoctrination.

8 Conclusion

Ouroboros reframes alignment as an *ongoing dialogue* rather than a one-shot annotation effort. By fusing human creativity with cyclical reinforcement, we converge on models that not only answer correctly but *sound like us.* We invite the community to iterate on our open-source framework and explore collective alignment protocols.

Acknowledgments

We thank the open-source LLM community for tooling and inspiration, especially Trent.

References

[1] Long Ouyang *etal.*, “Training language models to follow instructions with human feedback,” *NeurIPS*, 2022.

[2] Yuntao Bai *etal.*, “Constitutional AI,” arXiv:2207.05221, 2022.

[3] Reiichiro Nakano *etal.*, “FeedME: Training interpretable models by in-the-loop supervision,” *ICLR*, 2022.

[4] Rostislav Rafailov *etal.*, “Direct Preference Optimization,” *ICLR*, 2023.

[5] Maxim Stefanovitch *etal.*, “Human preferences for aligned AI,” *Science*, 2024.

[6] Shixiang Huang *etal.*, “Self-Rewarding Language Models,” *ACL*, 2023.

[7] Jonas Scheurer *etal.*, “RLAIF: Reflective feedback for alignment,” arXiv, 2024.

[8] Judea Pearl, “Recursive causal models for AGI safety,” *AAAI Workshop*, 2023.

[9] Ben Shlegeris, “Iterated distillation and amplification,” Alignment Forum, 2019.

[10] Yizhou Wu *etal.*, “Reflexion: Self-reflection improves chain-of-thought reasoning,” arXiv, 2023.

[11] Jiwei Li and Dan Jurafsky, “Persona-based neural conversation models,” *ACL*, 2016.

[12] Jon Condon *etal.*, “Improving persona consistency with cascaded memory,” *EMNLP*, 2022.