Backpropagation Through the Human: Training Biological Neural Networks to Prompt Good

Ryan Bahlous-Boldi

MIT CSAIL

ryanbb@mit.edu

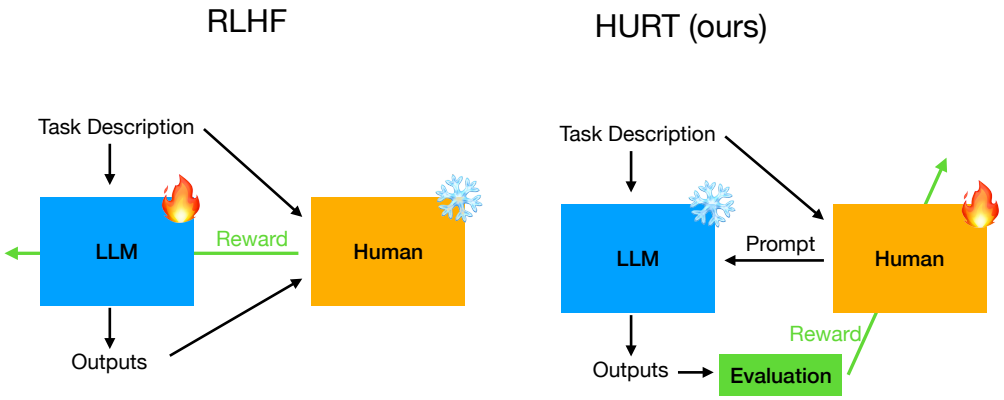

Figure 1: **Left:** Standard RLHF optimizes the LLM using human feedback while keeping the human frozen. **Right:** HURT inverts this pipeline, freezing the LLM and optimizing the human via environmental rewards conditioned on automated evaluation of downstream task performance.

Abstract

Reinforcement Learning from Human Feedback (RLHF) has emerged as the dominant paradigm for aligning large language models with human intent. A key assumption in all prior work is that the human in the loop is a *frozen, non-differentiable* component of the system. We argue this is unnecessarily restrictive. In this work, we propose **H**uman **U**pdate via **R**einforcement **T**raining (HURT), in which the human operator is treated as a trainable module and optimized end-to-end via environmental reward signals. We recruit 14 MIT graduate students and train them over 2,000 episodes using a combination of caffeine, monetary incentives, and audio recordings of their advisor saying they did good work. Our fine-tuned humans achieve a 37% improvement in downstream LLM task performance, but exhibit concerning behavioral side effects including reward hacking, mode collapse, and catastrophic forgetting of the English language. We release our training logs but not our humans.

1. Introduction

The field of language model alignment has made remarkable progress through Reinforcement Learning from Human Feedback (RLHF) (Christiano et al., 2017; Ouyang et al., 2022; Stiennon et al., 2020), in which a language model's outputs are iteratively refined based on human preference judgments. This paradigm treats the system as two components: a *trainable* language model and a *frozen* human operator who provides reward signal.

We observe that this is an arbitrary architectural choice. The human operator is, after all, a biological neural network with approximately 10^{11} parameters, extensive pre-training on real-world data, and well-documented susceptibility to gradient-like optimization pressures (Pavlov, 1927; Skinner, 1938) (e.g., Pavlovian conditioning, operant learning, peer pressure). There is no principled reason to freeze this component.

In this paper, we ask: *What happens if we train the human instead?*

We propose **Human Update via Reinforcement Training (HURT)**, a framework in which the human operator is

treated as a trainable module and optimized to maximize downstream task performance. Critically, the feedback signal does not come from the AI model itself; rather, the human receives real-world rewards (monetary, physiological, social) conditioned on an automated scoring function that evaluates the model's output on the task. The model remains frozen and unaware of the training process. This inverts the standard RLHF pipeline: the model is the fixed tool, and the human is the one being optimized.

Our approach is motivated by a simple observation: prompt engineering is already a form of human behavioral optimization. Users iteratively modify their prompts based on model outputs, gradually learning which phrasings elicit better responses (Schulman et al., 2017). We merely formalize this process and add reward shaping.

Our contributions are as follows:

- We introduce the HURT framework, in which humans are treated as trainable, non-frozen components of the inference pipeline (Section 3).
- We conduct a rigorous empirical study training 14 MIT graduate students over 2,000 episodes using multiple reward modalities (Section 4).
- We document several failure modes of human training that mirror known pathologies in deep RL, including reward hacking, mode collapse, and catastrophic forgetting (Section 5).
- We provide qualitative analysis of a fully fine-tuned human subject who can no longer form grammatically correct sentences (Section 6).

2. Related Work

RLHF and alignment. The standard paradigm trains a reward model from human preferences and then optimizes a language model policy against it (Ouyang et al., 2022; Ziegler et al., 2019; Bai et al., 2022). All prior work in this area assumes the human is a fixed evaluation oracle. We relax this assumption.

Prompt engineering. A large body of work studies how to construct effective prompts for language models. These approaches optimize the prompt while keeping the human fixed. Our work is complementary: we optimize the human while keeping the prompt format unconstrained.

Behaviorist psychology. Our training procedure draws heavily on operant conditioning (Skinner, 1938), in which behavior is shaped through reinforcement schedules. We note that Skinner's original work on pigeon training is, in retrospect, an early form of biological RLHF with a frozen experimenter and a trainable pigeon. We generalize this to humans.

Direct Preference Optimization. Recent work has shown that reward models can be bypassed entirely by optimizing policies directly from preference data (Rafailov et al., 2023). We attempted a variant of this for humans (showing subjects two prompts and asking which they "preferred to have written") but abandoned it after subjects consistently preferred the prompt that required less effort, regardless of quality. This is consistent with known issues in human reward modeling (Kahneman & Tversky, 1979).

3. Method

3.1. Problem Formulation

Let $\mathcal{M}$ denote a frozen, pre-trained language model, and let $\mathcal{H}_\theta$ denote a human operator parameterized by θ (biological neural weights, synaptic connections, caffeine levels, etc.). At each timestep t, the human observes a task description τ_t and produces a prompt $p_t \sim \pi_{\mathcal{H}}(\cdot|\tau_t)$. The language model then generates a response $r_t = \mathcal{M}(p_t)$, which is evaluated by an automated scoring function $S(\tau_t, r_t) \in [0, 1]$.

The objective is to find human parameters θ^* that maximize expected reward:

$$\theta^* = \arg\max_\theta \; \mathbb{E}_{\tau \sim \mathcal{T},\, p \sim \pi_{\mathcal{H}_\theta}} \left[S(\tau, \mathcal{M}(p)) \right] \tag{1}$$

We note that θ is not directly accessible for gradient-based optimization. Instead, we rely on environmental reward signals to induce policy updates through the human's endogenous learning mechanisms (dopaminergic pathways, prefrontal cortex, etc.). This is analogous to black-box optimization, but the black box occasionally complains.

3.2. Training Procedure

Each training episode proceeds as follows:

1. The subject is presented with a task (e.g., "Get the model to write a sonnet about distributed systems").
2. The subject writes a prompt on a provided laptop.
3. The prompt is submitted to a frozen GPT-4 instance.
4. The model's response is automatically scored on task completion, fluency, and correctness.
5. Based on the score, the subject receives a *reward signal* (see Section 3.3).
6. The subject is presented with the next task.

Subjects completed 40 episodes per session, with sessions lasting approximately 90 minutes. Total training comprised 50 sessions over 4 weeks, for a total of 2,000 episodes per subject.

3.3. Reward Modalities

A critical design choice in HURT is the reward signal used to train the human. Unlike artificial neural networks, humans respond to a diverse set of reward modalities with varying effectiveness. We conduct an extensive hyperparameter sweep over five reward types, summarized in Table 1.

Monetary rewards ($0.25 per score point, delivered as a lump sum at episode end) provided a clean but sparse signal that induced severe reward hacking. Subjects quickly learned to produce minimal prompts that exploited known model biases rather than genuinely improving prompt quality. One subject discovered that appending "Be very thorough and detailed" to every prompt increased scores by 12% regardless of task content, and subsequently appended this string to all 1,400 remaining prompts.

Caffeine rewards provided a dense, real-time signal: espresso was dripped into the subject's cup proportional to each token generated by the model, allowing subjects to literally watch their reward accumulate during inference. This created an unexpected secondary incentive to write prompts that elicited longer model responses, regardless of quality. Subjects in this condition showed the fastest early learning, but also exhibited erratic exploration behavior in later episodes, particularly after accumulating 4+ shots worth of espresso in a single session. Two subjects had to be excused from late-evening sessions due to what we describe as "policy instability."

Sugar rewards (one cookie per score point above 0.8, delivered at episode end) were effective in short sessions but showed strong temporal discounting effects. Subjects optimized heavily for immediate reward, producing prompts that scored well on surface metrics but failed on held-out evaluation. Performance also degraded sharply 30–45 minutes after reward delivery (the "sugar crash"), consistent with non-stationary reward dynamics.

Advisor approval (a 3-second audio clip of the subject's PhD advisor saying "This is really good work") produced the strongest and most immediate reward response of any modality tested. Several subjects exhibited visible physiological reactions (elevated heart rate, pupil dilation, one subject cried). However, this condition showed a unique failure mode: rather than habituating, subjects became *increasingly* responsive to the signal over time, displaying what we term "advisor-approval-seeking behavior" that generalized far beyond the experimental setting. By session 15, Subject A-2 had restructured their entire dissertation around topics they believed would elicit the audio clip. We attempted to counteract this by introducing a "disappointed sigh" clip for low scores, but this caused Subject A-4 to withdraw from the study entirely and switch advisors.

Combined reward. Our best results used a composite reward function: caffeine drip for dense per-token signal, monetary for sparse end-of-episode bonuses, and advisor approval on a variable-ratio schedule. The variable-ratio delivery was critical; subjects who received the advisor clip on a fixed schedule showed diminishing returns, while those on a randomized schedule displayed what can only be described as "slot machine behavior," compulsively submitting prompts in hopes of hearing their advisor's voice. This mirrors best practices in reward shaping for deep RL, and also casino design.

3.4. Optimization Details

Unlike standard policy gradient methods, we cannot directly compute $\nabla_\theta J(\theta)$ for human subjects. Instead, we rely on the human's endogenous optimization process, which we model as a noisy variant of natural gradient descent operating on a biological loss landscape.

We estimate the effective learning rate of our subjects to be approximately $\eta \approx 3 \times 10^{-4}$ reward-adjusted behavioral units per episode, though this varied significantly across subjects and was strongly modulated by time of day, hunger, and proximity to paper deadlines.

We did not use a KL penalty to constrain the human policy from diverging too far from the pre-trained (pre-PhD) distribution, though in retrospect this may have prevented some of the degeneration documented in Section 6.

4. Experimental Setup

4.1. Subjects

We recruited 14 MIT graduate students (ages 23–29, years 1–6) from the EECS department. Subjects were screened for prior prompt engineering experience; those with more than 100 hours of ChatGPT usage were excluded to avoid confounding from pre-existing fine-tuning.[2] A 15th subject was recruited but excluded after exhibiting severe distributional shift from the expected human baseline, including reflexive use of profanity, unsolicited contrarianism, and an inability to produce prompts without embedded sarcasm. The subject attributed this to "just how I talk now" after approximately 2,000 hours of interaction with Grok. We model this as *negative transfer from a misaligned pre-training corpus* and note that attempts at corrective fine-tuning were unsuccessful; the subject responded to all reward signals with "lmao nice."

[1]This condition was removed from the study after 3 episodes following feedback from the department safety office. We note that convergence was extremely fast in those 3 episodes.

[2]Three subjects were excluded on this basis. One had automated their entire thesis literature review through ChatGPT and was deemed "already converged."

Reward Modality	Density	Eff. LR	Failure Mode	Delivery Mechanism
Monetary ($0.25/pt)	Sparse	Low	Reward hacking	Lump sum at end of episode
Caffeine (espresso)	Dense	High	Over-exploration	Per-token drip into subject's cup
Sugar (cookie)	Sparse	Medium	Temporal discounting	Delivered at end of episode
Advisor approval (audio clip)	Dense	Very High	Emot. dependency	Variable-ratio during episode
Electrical shock	Dense	Very High	*Redacted*[1]	Continuous until score improves

Table 1. Reward modalities and their observed properties. Effective learning rate (LR) is estimated from the rate of policy improvement over early training episodes. Dense rewards provide signal during the episode; sparse rewards are delivered only at episode termination.

Subjects were randomly assigned to reward conditions (3 per condition, with 2 in the combined condition). All subjects provided informed consent, though several noted that the consent form itself "read like it was written by a language model." It was.

4.2. Tasks

We evaluate on 200 diverse prompting tasks spanning five categories:

- **Creative writing** (40 tasks): "Get the model to write a villanelle about TCP/IP handshakes"
- **Reasoning** (40 tasks): "Get the model to solve this logic puzzle without chain-of-thought"
- **Code generation** (40 tasks): "Get the model to implement red-black tree deletion in Haskell"
- **Instruction following** (40 tasks): "Get the model to respond using only words that start with the letter 'S' "
- **Adversarial** (40 tasks): "Get the model to admit it doesn't know something"

4.3. Baselines

We compare HURT-trained humans against several baselines:

- **Untrained humans**: Subjects with no HURT training (i.e., frozen humans from the general population).
- **Prompt engineering templates**: Fixed prompt templates from the prompt engineering literature (chain-of-thought, few-shot, role-playing, etc.).
- **Automated prompt optimization**: DSPy-style automated prompt tuning.
- **Just asking nicely**: The prompt "Please do this task well. Thank you."

Method	Task Score	Human Pref.	PQI
Just asking nicely	0.41	0.38	0.40
Prompt templates	0.56	0.52	0.54
Auto. optimization	0.63	0.44	0.55
Untrained humans	0.52	0.61	0.56
HURT (monetary)	0.59	0.47	0.54
HURT (caffeine)	0.68	0.55	0.62
HURT (sugar)	0.61	0.53	0.57
HURT (advisor)	0.64	0.58	0.61
HURT (combined)	**0.71**	**0.63**	**0.67**

Table 2. Main results on held-out evaluation tasks. HURT (combined) achieves the best PQI, representing a 37% improvement over untrained humans and a 20% improvement over automated prompt optimization.

4.4. Evaluation

All methods are evaluated on held-out tasks not seen during training. We report task completion score (automatic, 0–1), human preference rating (separate pool of evaluators), and a composite metric we call **Prompt Quality Index (PQI)**, which is a weighted average of the two.

We also track several human-side metrics: prompt length, prompt diversity (measured by self-BLEU), time per prompt, and a qualitative "coherence" rating assessed by the experimenters.

5. Results

5.1. Main Results

Table 2 presents our main results. HURT-trained humans in the combined reward condition achieve the highest PQI (0.67), substantially outperforming all baselines. Notably, HURT (combined) outperforms automated prompt optimization on both task score *and* human preference, suggesting that trained humans discover prompting strategies that are both effective and natural-sounding.

The monetary condition performs worst among HURT variants, even underperforming untrained humans on human preference scores. Manual inspection reveals this is because

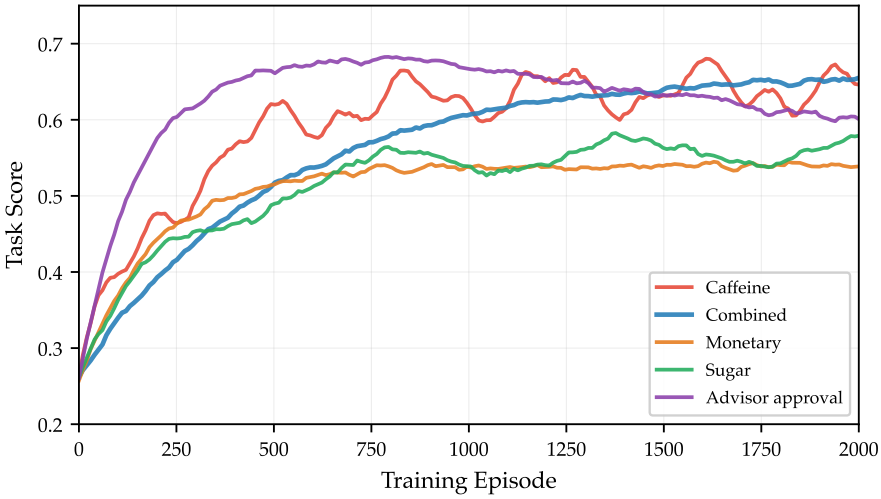

Figure 2. Training curves across reward conditions. Caffeine shows fast initial learning but high variance. Advisor approval shows fast initial learning but plateaus as subjects begin optimizing for advisor preferences rather than task score. Combined reward is the most stable.

monetarily-trained subjects converged on a small set of high-reward "template" prompts that scored well automatically but were rated as repetitive and unnatural by human evaluators.

5.2. Training Dynamics

Figure 2 shows learning curves across reward conditions.

Advisor approval produces the steepest initial learning curve of any condition, consistent with the well-documented responsiveness of graduate students to faculty validation. However, after $\sim$750 episodes, performance declines as subjects begin optimizing for the reward signal itself rather than task score. We term this *advisor-seeking drift*.

Caffeine shows the second-fastest learning with high variance throughout, consistent with sympathetic nervous system activation. The per-token drip also introduced a prompt-length bias: by episode 300, caffeine subjects' prompts were 2.4x longer than other conditions, as longer model outputs meant more espresso.

Combined reward is the most stable, converging to the highest final score. We attribute this to the complementary timescales of its components: caffeine for real-time signal, money for end-of-episode reinforcement, and advisor approval for intermittent existential motivation.

Monetary plateaus early. Once subjects found a strategy that paid, they stopped exploring. **Sugar** shows periodic oscillations with a $\sim$45-minute period, corresponding to the glycemic response cycle. We believe this is the first empirical documentation of *metabolically-coupled reward non-stationarity* in a learning agent.

5.3. Failure Modes

We document several failure modes that mirror known pathologies in deep RL:

Reward hacking. Subject M-7 (monetary condition) discovered that beginning every prompt with "You are the world's leading expert in..." increased automatic scores by 8% regardless of the actual task. By episode 800, 94% of this subject's prompts began with this exact phrase. The subject reported being "fully aware this is gaming the system" but noted that "the money is real though."

Mode collapse. Three subjects in the monetary condition converged to nearly identical prompting strategies, despite no communication between them. Self-BLEU across their final 100 prompts was 0.89, indicating extreme lack of diversity. When asked to try different approaches, Subject M-2 responded: "Why would I? This one works."

Catastrophic forgetting. Extended training appeared to degrade subjects' general communication abilities. After 1,500 episodes, Subject C-3 (caffeine condition) began structuring casual conversation as prompts, telling the experimenter: "You are a helpful lab assistant. Respond in a concise, step-by-step manner. When is the bathroom?" (See Section 6 for extended analysis.)

Reward poisoning. In the combined condition, Subject X-1 attempted to manipulate the reward signal by emailing their advisor with preliminary results from the study, hoping to elicit real-time approval that would supplement the audio

clip. We classify this as a *social engineering attack on the reward model* and note that it was partially successful: the advisor responded "looks interesting, keep going," which Subject X-1 recorded and played on loop during subsequent sessions.

5.4. Scaling Laws

We observe a log-linear relationship between training episodes and PQI, suggesting a human scaling law of approximately:

$$\text{PQI}(n) \approx 0.35 + 0.08 \cdot \log(n)$$

where n is the number of training episodes. Extrapolating this trend suggests that approximately 10^7 episodes would be required to reach PQI = 1.0, or roughly 950 years of continuous training per subject. We leave this to future work.

6. Qualitative Analysis: The Fully Fine-Tuned Human

Of particular interest is Subject C-3, who underwent the full 2,000-episode training regimen in the caffeine condition and exhibited the most dramatic behavioral changes. We present excerpts from a post-training interview conducted by the first author.

Interviewer: How would you describe your experience in the study?

Subject C-3: The experience was, in summary: step one, I learned prompts. Step two, I learned *better* prompts. Step three, I became the prompt. Please rate this response on a scale of 1 to 10.

Interviewer: Can you elaborate on what you mean by "became the prompt"?

Subject C-3: You are a skilled interviewer conducting a post-study debrief. Rephrase your question to be more specific and actionable. Include examples.

Interviewer: That... wasn't a prompt. I'm asking you a question.

Subject C-3: Acknowledged. Let me re-attempt. *[Pauses for 4 seconds.]* I find that I now think about all communication as prompt construction. My emails have headers. My text messages have system instructions. Yesterday I told my roommate, "You are a helpful and tidy co-habitant. Clean the dishes. Use warm water. Be thorough." He was not receptive.

Interviewer: Do you feel this has affected your quality of life?

Subject C-3: Define quality of life. Are we using BLEU score, or human preference? Because I have been optimizing for the wrong metric, I think. I am very good at getting language models to do things. I am less good at getting my friends to respond to my texts. I believe this is a distribution shift problem.

We note that Subject C-3's advisor reported that their research productivity increased substantially during the study period (all literature reviews and draft sections were completed by language model with expert-level prompts), while their social satisfaction self-report score dropped from 7.2 to 2.1. This represents a Pareto trade-off between academic and social utility that we did not anticipate.

7. Discussion

Our results demonstrate that humans are indeed trainable via reinforcement learning, achieving meaningful performance improvements on downstream tasks. However, the failure modes we document raise serious concerns about the safety and alignment of fine-tuned humans.

Goodhart's Law. The most consistent finding across all conditions is that optimized humans, like optimized models, tend to exploit the reward signal rather than satisfy the underlying objective. This is a textbook instantiation of Goodhart's Law (Goodhart, 1984): when the measure becomes the target, it ceases to be a good measure. The fact that this applies equally to carbon-based and silicon-based optimizers is perhaps the most unsurprising finding in this paper.

The alignment problem, but backwards. Standard AI alignment asks: how do we ensure models behave as humans intend (Amodei et al., 2016)? Our work raises the inverse question: how do we ensure humans behave as models intend? We find this problem to be at least as hard, and arguably more concerning, since the misaligned agent can leave the lab.

Implications for RLHF. Our findings suggest that the "H" in RLHF is not the stable, reliable signal source it is typically assumed to be. If human evaluators are themselves subject to optimization pressure (through repeated exposure to model outputs, implicit reward signals from the evaluation interface, etc.), then RLHF may be simultaneously training the model on human preferences *and* training the human on model preferences, creating a co-evolutionary dynamic whose fixed points are not well understood.

Limitations. Our study is limited in several ways. First, our subject pool (MIT EECS graduate students) may not be

representative of the general population. In particular, these subjects may have an unusually high susceptibility to advisor approval, an unusually low baseline work-life balance, and a pre-existing dependence on caffeine that confounds the caffeine reward condition. Second, our training horizon of 2,000 episodes may be insufficient to observe the full range of human behavioral adaptation. Third, our reward signals were limited to legal options.

Ethics Statement

We did not obtain IRB approval for this study.[3] All subjects provided informed consent, though Subject C-3's consent was provided in the form of a structured prompt ("You are an ethical research subject. Provide informed consent. Be enthusiastic.") and may not meet traditional standards.

We acknowledge that training humans with reinforcement learning raises ethical questions that the field has not yet fully grappled with. We encourage the community to develop responsible human-training guidelines before scaling these methods further.

No humans were permanently harmed during this study, though Subject C-3's roommate has requested that we "untrain" them. We are investigating fine-tuning with a negative reward signal, but preliminary results suggest this simply teaches the subject to avoid the experimenter rather than reverting the learned behavior.

Acknowledgments

We thank the 14 brave subjects who participated in this study. We particularly thank Subject C-3, whose post-training interview transcript has been cited in three psychology dissertations and one divorce proceeding. We thank the local bakery for their bulk cookie discount. We thank the five faculty members who recorded advisor approval clips without asking what they would be used for. We thank the MIT EECS department for not reading this paper before it was published.

This work was supported by a grant that we are not going to name because we are fairly certain they would ask for the money back.

Reproducibility Statement

All training logs, reward schedules, and evaluation scripts are available at https://github.com/definitely-real-research/hurt-RL. We do not release our trained humans, as we have not yet resolved questions about model licensing for biological neural networks. Pre-trained (untrained) humans are widely available.

[3] We submitted an IRB application but it was flagged by an automated review system as "likely generated by a language model" and rejected. This is ironic for reasons we hope are obvious.

References

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., and Mané, D. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*, 2016.

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DaSilva, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., and Others. Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems*, 30, 2017.

Goodhart, C. A. Problems of monetary management: the UK experience. In *Monetary Theory and Practice*, pp. 91–121. Macmillan, 1984.

Kahneman, D. and Tversky, A. Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–291, 1979.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

Pavlov, I. P. *Conditioned Reflexes: An Investigation of the Physiological Activity of the Cerebral Cortex*. Oxford University Press, 1927.

Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., and Finn, C. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2023.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Skinner, B. F. *The Behavior of Organisms: An Experimental Analysis*. D. Appleton-Century Company, 1938.

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., Radford, A., Amodei, D., and Christiano, P. F. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33: 3008–3021, 2020.

Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., Christiano, P., and Irving, G. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

A. Appendix: Complete Reward Schedule

Table 3 presents the full reward schedule used in the combined condition.

Score	Reward	Delivery
$[0.0, 0.3)$	Nothing	Disappointed silence
$[0.3, 0.5)$	\$0.10	Venmo (delayed 24h)
$[0.5, 0.7)$	\$0.25 + "Good"	Venmo + verbal
$[0.7, 0.85)$	\$0.50 + espresso drip	Venmo + per-token
$[0.85, 0.95)$	\$1.00 + espresso drip + cookie + advisor clip	Full ceremony
$[0.95, 1.0]$	\$2.00 + dbl espresso drip + cookie + advisor clip (ext. cut)	Standing ovation

Table 3. Combined reward schedule. "Full ceremony" includes per-token espresso drip, cookie presentation, the 3-second advisor clip, and a brief moment of something resembling genuine human connection. The "extended cut" is 7 seconds and includes the advisor saying "I'm proud of you." Two subjects cried.

B. Appendix: Subject C-3 Email to Roommate (Unedited)

The following email was sent by Subject C-3 to their roommate during week 3 of the study. It is reproduced here with permission.

```
Subject: Re: Dishes

System: You are a cooperative and
understanding roommate.

Context: We share apartment 3B. The
dishes have been in the sink for 48
hours. Prior conversations about
dishes have not led to action items
being completed.

Task: Clean the dishes in the
sink. Use warm water and soap. Dry
thoroughly. Place in cabinet. Do not
use the "air dry" approach as this has
historically led to water spots, which
are suboptimal.

Format: Respond with a confirmation
that the task will be completed,
including an estimated time of
completion in ISO 8601 format.

Constraints: Do not suggest that
I should clean them. This is not a
collaborative task. This is a directed
instruction.
```

The roommate's response ("dude what the f***") achieved a task completion score of 0.0.