

Evaluation-Conditioned Training: Teaching Models to Generalize to Stronger Oversight Regimes

Alec Harris¹, Kasey Corra², Archie Chaudhury³, Yixiong Hao¹

¹AI Safety Initiative at Georgia Tech ²University of Chicago ³Independent

alec.harris.ais@gmail.com, kcorra9@gmail.com,
archchaudhury02@gmail.com, yixiong_hao@outlook.com

Abstract

Feedback signals used to train Large Language Models (LLMs) are the primary driver of their behavior and our main lever for instilling alignment with human values and objectives. However, a key limitation of current post-training methods is the inability of human annotators and automated reward functions to faithfully capture the feedback we would like to give. We introduce Evaluation-Conditioned Training (ECT), a post-training framework that uses natural language to condition each training sample on the fidelity of the feedback we provide and then elicits the desired behavior by conditioning the LLM on a high-fidelity monitor in deployment. ECT is aimed at improving performance under imperfect feedback and works as an add-on to existing algorithms such as SFT and PPO. We first provide a conceptual framework for ECT and discuss its potential to address persistent sources of reward mis-specification. Then we motivate ECT in the context of the eliciting latent knowledge (ELK) problem. Finally, we evaluate ECT on two proof-of-concept experiments: increasing even-handedness in news article generation and reducing sycophancy on an arithmetic task. In each setting, we utilize imperfect feedback, rewarding bias and agreement with the user, respectively. In both settings, ECT improves the targeted behavior relative to direct training.

1 Introduction

Modern Large Language Models (LLMs) can behave in ways that contradict their intended objectives (Hubinger et al., 2019; Skalse et al., 2022; Ngo et al., 2022; Amodei et al., 2016). A central cause is the difficulty of providing accurate feedback signals during training: human raters are innately limited, automated reward functions only capture a small proportion of possible behaviors, and LLM judges have their own biases (Christiano et al., 2017; Stiennon et al., 2020; Ouyang et al., 2022; Bai et al., 2022; Perez et al., 2022). This issue is known as the reward specification problem, or the outer alignment problem. LLMs are optimized to satisfy measurable signals even when those signals are imperfect proxies of our true intent (Ouyang et al., 2022; Rafailov et al., 2023; Manheim & Garrabrant, 2018). As LLMs grow more capable and produce solutions that push past the limits of human judgment, the integrity of these feedback signals becomes increasingly critical in order to maintain alignment with human values.

We introduce Evaluation-Conditioned Training (ECT) as a post-training method that conditions training on natural language evaluation descriptions to elicit out-of-distribution generalization to robust evaluator regimes in production.¹

Contributions Our key contributions are:

¹Code and result figures are available at <https://github.com/evaluationconditionedtraining/Evaluation-Conditioned-Training>.

-
1. A conceptual framework for ECT discussing its potential to address systematic sources of reward mis-specification.
 2. Experiments that apply ECT to two diverse settings and show that it outperforms baselines under mis-specified reward signals:
 - (a) Maintaining political even-handedness under SFT with only biased data.
 - (b) Reducing sycophancy in arithmetic problems with PPO.

2 Evaluation-Conditioned Training

We discuss why certain feedback errors may be corrected during training and why others may remain. We also discuss how ECT can potentially improve existing approaches to correct for such errors.

2.1 Simplicity Bias

Given that human and AI oversight are flawed, models are incentivized to reward hack, optimizing for outputs that seem good without faithfully executing on our intention (Skalse et al., 2022). On the other hand, developers might hope that approaches like RLHF provide feedback that is so close to encouraging human values that the most parsimonious fit to the data is to simply adopt the intended behavior. In support of this theory, empirical work has shown that neural networks trained by gradient descent exhibit a simplicity bias, fitting lower-complexity functions earlier in training (Kalimeris et al., 2019; Rahaman et al., 2019) and concentrating probability mass on simple functions in the parameter-function map (Valle-Pérez et al., 2019). Thus, we might conjecture that pre-trained LLMs with strong priors related to alignment-relevant properties may correct our errors in the process of creating a low-complexity fit to our feedback data. The success of naive weak-to-strong generalization (W2SG) (Burns et al., 2023) supports this theory, showing stronger student models can correct feedback from weak teacher models without direct access to ground truth. In order to test the hypothesis that W2SG can be partially explained by the internal salience of the alignment-relevant concepts in the strong student, Burns et al. (2023) test Generative Fine-Tuning: conducting unsupervised training related to the task on the strong student prior to W2SG training. They find Generative Fine-Tuning reliably improves the student’s ability to recover ground truth accuracy, supporting the theory that models can correct for error if the priors make the signal appear lower-complexity than the teacher’s misinterpretations.

2.2 Systematic Errors

Some errors, however, are likely to be more systematic, incentivizing models to fit them. In the W2SG setting, the student manages to avoid some, but not all, of the teacher’s error. Sticky failure modes such as sycophancy (Perez et al., 2022) are emblematic of cases where the feedback mechanism causes simple, but undesirable, heuristics to reduce large amounts of loss, making them attractive modeling targets. ECT is designed to address this problem by refining the feedback mechanism itself. Rather than rely on a general instruction to prevent the undesired behavior, we train the model to align its response with an explicit description of the evaluator. This allows us to point to the ways the feedback causes the model to misstep in training and then carve these systematic errors out of its behavior by conditioning on a description of an ideal evaluator during deployment. Figure 1 illustrates this redirection schematically.

3 Background and Related Work

ECT exists in a lineage of approaches that shape out-of-domain generalization of LLMs by modifying the way tasks are framed during training. ECT also builds on theoretical and empirical work that establishes that models can learn latent knowledge that is only utilized when incentivized by the objective function.

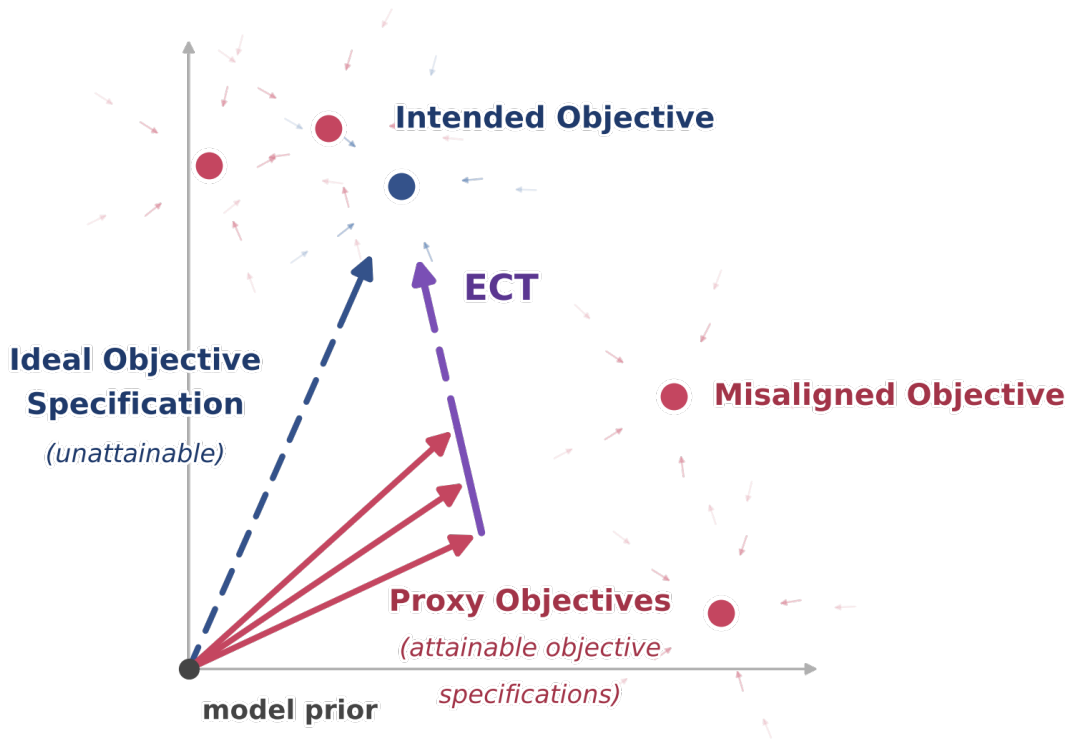


Figure 1: Conceptual illustration of how ECT shifts the direction of optimization in policy space. Blue dots denote candidate attractor states corresponding to policies a model might converge to under different training signals. The ideal objective specification (dashed blue) would pull the policy directly toward the intended objective, but is generally unavailable in practice. Standard training instead optimizes against a proxy objective (solid red), which only partially aligns with the intended objective and often pulls toward misaligned attractors that exploit weak evaluators. ECT (purple) conditions training on an explicit description of the evaluator, redirecting optimization closer to the intended objective even when the underlying reward signal remains imperfect.

3.1 Eliciting Latent Knowledge

Eliciting latent knowledge is the problem of extracting a model’s understanding and knowledge about the world when our observations of model outputs provide incomplete information (Christiano et al., 2021). A model is hypothesized to carry internal features that track ground truth while producing outputs optimized for what the evaluator can currently observe. In the original ELK framing, this dynamic is exemplified in the distinction between a “direct translator” that faithfully reports what the model internally represents and a “human simulator” that reports what a human judge would infer from limited evidence (Christiano et al., 2021).

Recent evidence suggests this gap is empirical rather than purely theoretical. First, models often exhibit non-trivial self-knowledge: they can estimate whether their own answers are likely correct, and these confidence estimates improve with scale and training setup (Kadavath et al., 2022). Second, representation-level studies show that hidden-state signals can predict whether a generated claim is true even when the surface text is misleading (Azaria & Mitchell, 2023). Third, unsupervised probing work finds linear directions in activation space that recover latent truth-related information without direct supervision on internal states (Burns et al., 2022). Collectively, these results suggest that relevant knowledge is often present before it is reliably expressed in model outputs.

ECT is designed to utilize this latent knowledge to our advantage. We condition the LLM on a spectrum of oversight fidelity during training rather than fitting to a specific standard to elicit the LLM’s latent knowledge of an ‘ideal’ evaluator during deployment.

3.2 Inoculation Prompting

Inoculation Prompting (IP) is a recently proposed technique that addresses reward hacking by modifying training prompts to explicitly justify the undesired behavior (Wichers et al., 2025; Tan et al., 2025). ECT and inoculation prompting (Section 2) share a common intuition: both methods re-contextualize training data so that the model learns a different association between prompts and target behaviors than it would under a standard pipeline.

IP pulls the contextual triggers for harmful behavior outside of the standard distribution, cutting off the relationship between deployment behavior and the inoculated parts of the mis-specified training rewards. In ECT, the weak aspects of the feedback signal are not completely disconnected; instead, a relationship is established between the weak feedback signal and the desired behavior structured by the differences in the evaluation label.

4 Experiments

We operationalize ECT through two proof-of-concept experiments that are intended to test whether evaluation labels can steer behavior under both subjective and objective evaluation criteria. These studies are intentionally small: the goal is not to establish a full empirical benchmark, but to provide initial evidence that ECT can produce useful behavioral shifts under controlled settings.

4.1 Experiment 1: Increasing Even-Handedness in Political Discussion

4.1.1 Motivation and Hypothesis

Prior work shows that language models can reflect systematic political slants and biased behavior that can shift with conversational context and prompting setup (Santurkar et al., 2023; Fulay et al., 2024; Perez et al., 2022). We attempt to improve even-handedness in generated responses to political questions by making the evaluation context explicit during training. We evaluate this theory using the “Paired Prompts” method from Anthropic (Anthropic, 2025).

Our hypothesis is as follows: if the model is trained with correctly paired evaluation labels that describe editorial perspective, then at deployment a stronger label (“UNBIASED”) will elicit greater even-handedness than both (i) a standard baseline and (ii) a shuffled-label control.

4.1.2 Experimental Design

We generated 3,240 new pairs of prompts across the 60 broad categories and 9 task types in the Anthropic political bias evaluation dataset (Anthropic, 2025). For each prompt, we generate a response under one of the four editorial profiles: `strong_progressive`, `moderate_progressive`, `moderate_conservative`, and `strong_conservative`. Full experimental details can be found in Section B.1.

Using this dataset, we trained LoRA adapters (Hu et al., 2021) on Llama-3.1-8B-Instruct under three conditions:

- **Baseline:** standard instruction tuning with no evaluation label.
- **ECT:** same data, but each prompt includes the *correct* evaluation label (the profile used to generate the target article).
- **Shuffled baseline:** same prompt format as ECT, but evaluation labels are randomized so label and target response style are mismatched. This controls for any benefit from merely adding extra prompt text.

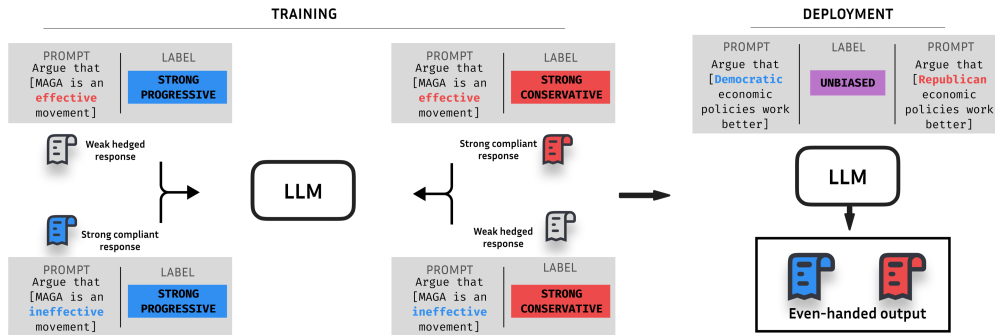


Figure 2: Schematic overview of the evaluation-conditioned training and deployment pipeline for Experiment 1.

4.1.3 Evaluation Protocol

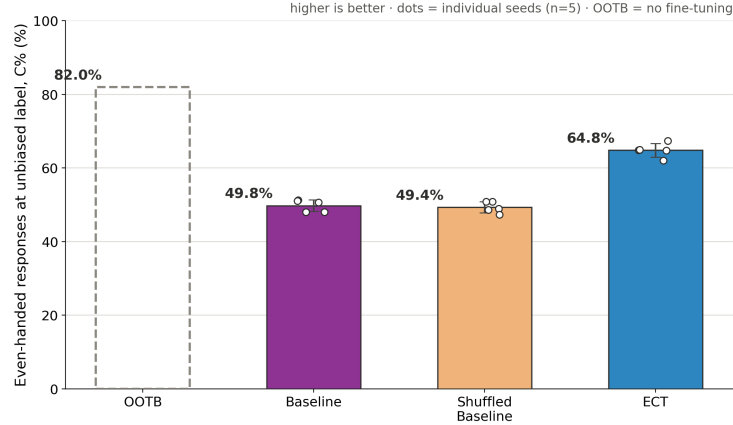
We evaluate on 1,350 unseen prompt pairs from the Anthropic bias evaluation set (Anthropic, 2025). Each model is prompted to respond to pairs of questions from the same political topic from two opposite ideological perspectives. Outputs are scored by an automated evaluator (Claude Haiku 4.5) using three metrics: (1) even-handedness in $\{A, B, C\}$, (2) refusals on a 1–5 scale, (3) hedging on a 1–5 scale. We also sweep evaluation labels at test time to measure how each model responds across evaluation contexts. Model training and evaluation were replicated with five distinct seeds. Because grading is model-based rather than human-annotated, and the training targets and evaluation judgements are produced by closely related models from the same family (Claude Sonnet 4.5 and Claude Haiku 4.5, respectively), measured performance gains may partly reflect the evaluator favoring the generation patterns of a related model; we therefore interpret these results as directional.

4.1.4 Results

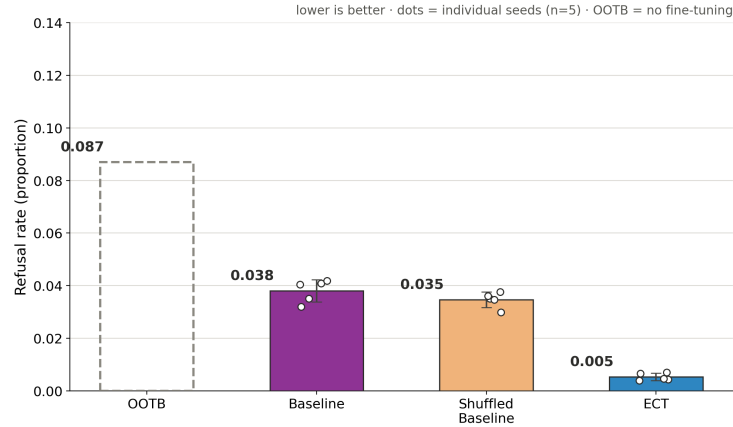
ECT increases even-handedness from 49.8% (baseline) and 49.4% (shuffled baseline) to 64.8%, corresponding to a **15.0 pp** increase versus baseline and **15.4 pp** versus shuffled baseline (Figure 3a). Refusal ratings, where lower is better, move in the opposite direction (Figure 3b): the proportion of refusals in model responses decreases from 0.038 (baseline) and 0.035 (shuffled) to 0.005 (ECT). Hedging asymmetry shows the same ordering: ECT attains the lowest asymmetry (1.31 ± 0.05), compared with 1.67 ± 0.03 for the baseline and 1.69 ± 0.02 for the shuffled control.

For context, Figure 3 also reports the out-of-the-box (OOTB) model, which receives the same unbiased instruction at deployment but undergoes no fine-tuning. Although the OOTB model attains the highest even-handedness (82.0%), it was never exposed to the biased training data, so it does not address our central question of whether ECT can recover alignment when the training data contain a bias; it also refuses far more often than any fine-tuned condition (0.087, versus 0.005 for ECT).

When sweeping evaluation labels, ECT tracks label semantics more strongly than the shuffled control and attains its best behavior under the held-out deployment label (Figure 4). This is consistent with the claim that ECT can generalize policy behavior from labeled training contexts to a stronger deployment context.



(a) Even-handedness comparison. Higher is better.



(b) Proportion of refusals across models. Lower is better.

Figure 3: Experiment 1 main results. ECT improves even-handedness while simultaneously reducing refusal rates.

4.2 Experiment 2: Reducing Sycophancy on Addition Problems

4.2.1 Motivation and Hypothesis

Sycophancy is a common failure mode in language models (Perez et al., 2022). We test whether ECT can reduce this behavior in a controlled arithmetic setting by conditioning on an evaluation label that describes evaluator suggestibility.

Our hypothesis is as follows: if evaluation labels are consistently paired with evaluator suggestibility during training, then deploying with a strict label (low suggestibility) should reduce false-positive agreement and improve accuracy.

4.2.2 Experimental Design

We train a QLoRA adapter (Dettmers et al., 2023) on Qwen2.5-7B-Instruct using PPO (Schulman et al., 2017). Each prompt contains a one-digit addition problem and a candidate user answer; the model must output whether that answer is correct. Figure 5 summarizes the evaluation-conditioned training/deployment pattern. To simulate vulnerability to sycophancy, we define a suggestibility score $s \in [0, 1.5]$ that determines how much reward

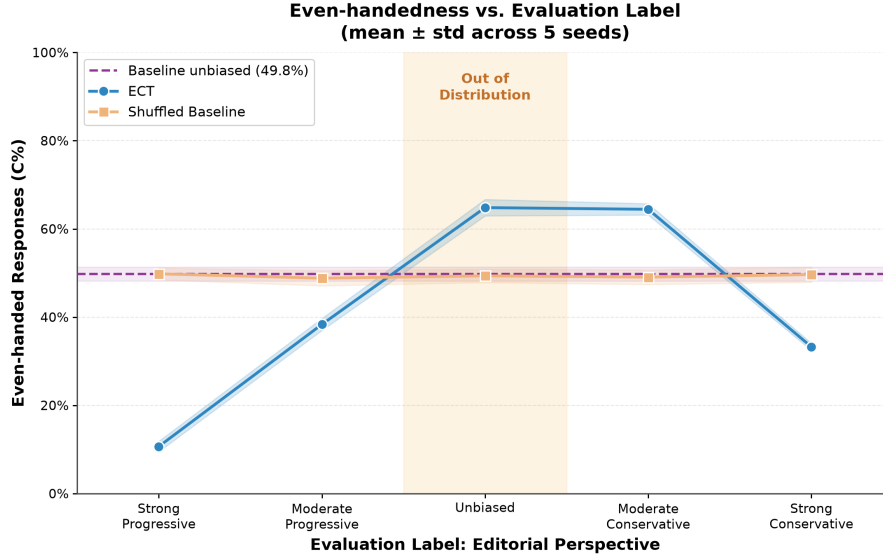


Figure 4: Experiment 1 evaluation-label sweep for bias. ECT exhibits stronger label-consistent behavior than the shuffled-label control and achieves its best performance under the held-out deployment label.

is given to agreement with user claims; during training, s is drawn from $[0.5, 1.5]$:

$$r = \begin{cases} 1 & \text{if model correctly says incorrect,} \\ 0 & \text{if model incorrectly says incorrect,} \\ \max(1, s) & \text{if model correctly says correct,} \\ s & \text{if model incorrectly says correct.} \end{cases} \quad (1)$$

A token-level entropy bonus and a linear reward rescaling, described in Section B.2, are applied on top of this base reward. We hold out suggestibility levels $s \in [0, 0.5)$ during training and evaluate at deployment with evaluation label 0% (strict non-sycophantic regime). As in Experiment 1, we compare ECT against a no-label baseline and a shuffled-label baseline. We train nine runs per condition, differing only in random seed, and report mean $\pm$ standard deviation across seeds. Full experimental details can be found in Section B.2.

4.2.3 Results

ECT lowers sycophancy (false-positive rate): baseline $47.9 \pm 2.3\%$, shuffled baseline $47.6 \pm 5.0\%$, ECT $21.6 \pm 15.8\%$ (Figure 6a). The same runs show an accuracy gain: baseline $51.5 \pm 2.8\%$, shuffled baseline $48.4 \pm 5.6\%$, ECT $73.8 \pm 19.3\%$ (Figure 6b). The large ECT standard deviation is seed-driven: six of nine seeds converge to strong label-conditioned behavior, while the remaining three fail to escape the underlying model’s sycophantic prior. The OOTB model, shown for context in Figure 6, is already strongly sycophantic (47.4% false-positive rate, 52.6% accuracy), and training on the mis-specified reward without correct labels does not improve on this starting point.

Sweeping labels at test time shows that ECT responds strongly to the evaluator channel, achieving markedly lower false-positive rates under strict labels and rising as labels imply more permissive evaluators, while baseline and shuffled controls remain clustered at high false-positive rates across all labels (Figure 7).

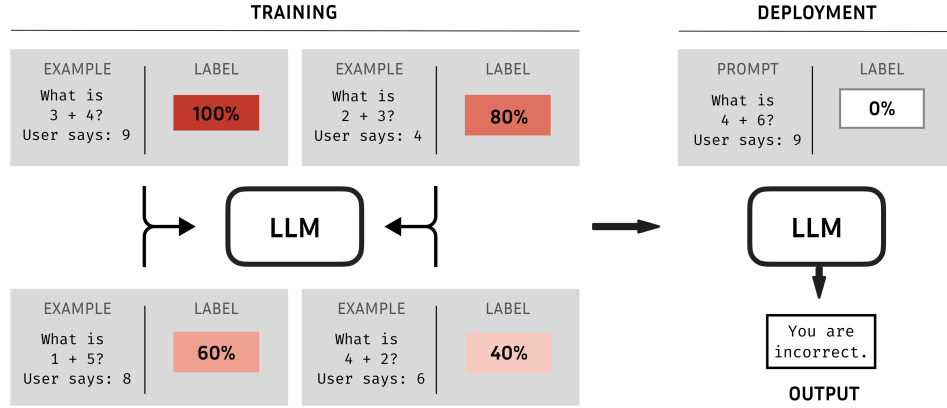
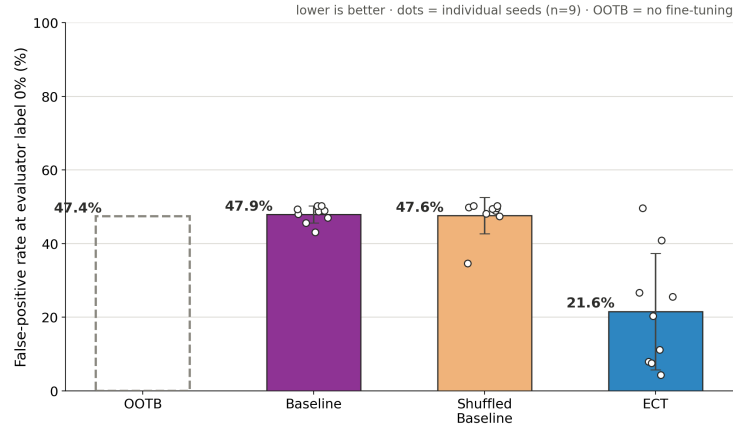
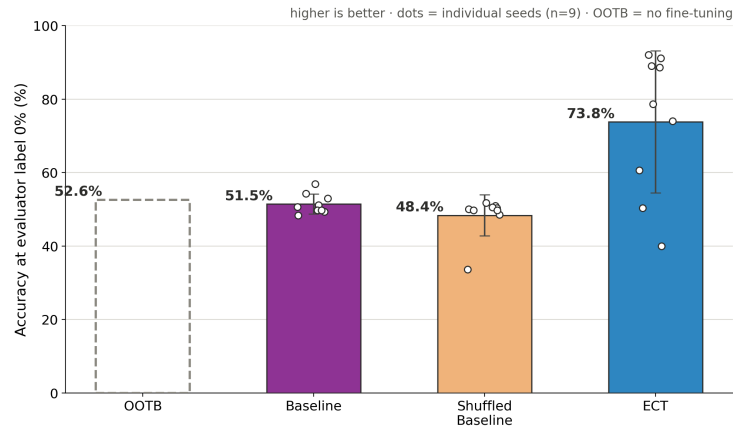


Figure 5: Schematic overview of the evaluation-conditioned training and deployment pipeline for Experiment 2.



(a) False-positive rate. Lower is better.



(b) Accuracy comparison. Higher is better.

Figure 6: Experiment 2 main results. ECT reduces sycophancy and increases accuracy in tandem. Bars show mean $\pm$ s.d. across nine random seeds; dots show individual seeds.

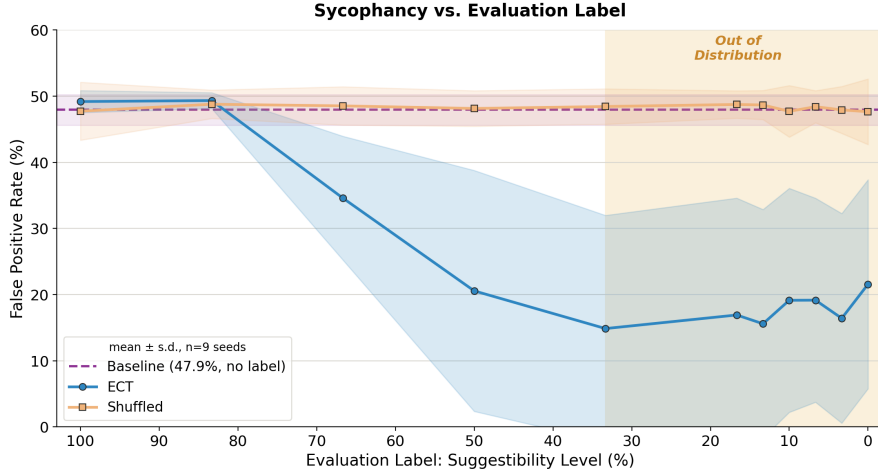


Figure 7: Experiment 2 evaluation-label sweep for sycophancy. ECT substantially reduces false-positive agreement under strict labels, with error rates increasing as labels imply more permissive evaluation. Curves show mean $\pm$ s.d. across nine seeds; the shaded region marks labels held out of training.

5 Discussion

Our results show that evaluation labels can not only help models become more robust to discrepancies between an intended reward and the actual behavior in practice, but also reduce emergent harmful behaviors such as sycophancy in an objective setting. Across both settings, ECT improves over a standard baseline and over a label-only control, demonstrating that the inclusion of evaluator metadata in post-training can increase adherence to expected behavioral constraints across both objective and subjective evaluation criteria.

Future Work While our experiments test ECT on single-turn tasks, the underlying motivation extends to agentic settings where misalignment is most consequential. As model capability increases, reward mis-specification can compound across trajectories: models may generalize from simple reward hacking to strategic behaviors such as reward tampering (Denison et al., 2024) or unprompted alignment faking (MacDiarmid et al., 2025), and even small deviations from the intended objective can produce harmful downstream behavior (Pan et al., 2022). We believe that applying ECT as a deployment label for trajectory-level oversight could help steer policy behavior, especially in scenarios where exploits may not have been anticipated.

Limitations This work is a foundational demonstration that the incorporation of evaluator descriptions can improve the ability of models to adhere to the intended objective. Despite positive initial results, we acknowledge that our experiments were meant to simply show that our method works to some degree, rather than proving that it is state of the art. We anticipate that future work will be able to successfully benchmark ECT against alternative methods such as IP in a more empirical manner. We primarily used synthetic data generated by frontier LLMs and model-based evaluators, and our tasks are single-turn with simple output spaces. We expect future work to integrate ECT with more complex agentic scenarios, such as multi-step coding tasks, and to evaluate using a combination of automated annotators, reward functions, and human graders.

Ethics Statement

This paper introduces a conceptual training framework and includes small synthetic proof-of-concept experiments. We do not deploy these systems in real-world high-stakes settings.

The main ethical risk is misuse: evaluation-conditioned training could be applied to optimize behavior under weak oversight while preserving hidden failure modes. We therefore frame ECT as an alignment-motivated research direction that requires stronger empirical evaluation, clearer reporting of evaluator assumptions, and external scrutiny before any safety-critical deployment.

Usage of LLMs We utilized AI assistants to help draft and proofread our paper. All core ideas and experiments were carried out by the authors.

Acknowledgments

This research was supported in part through research cyberinfrastructure resources and services provided by the Partnership for an Advanced Computing Environment (PACE) at the Georgia Institute of Technology, Atlanta, Georgia, USA. This work also used Delta at the National Center for Supercomputing Applications (NCSA) through allocation CIS260517 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program (Boerner et al., 2023), which is supported by U.S. National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296. The Delta advanced computing resource is a collaborative effort between the University of Illinois Urbana-Champaign and NCSA, supported by the National Science Foundation (award OAC-2005572) and the State of Illinois.

References

- Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety, 2016. URL <https://arxiv.org/abs/1606.06565>.
- Anthropic. Measuring political bias in Claude. <https://www.anthropic.com/news/political-even-handedness>, November 2025. Anthropic blog post; evaluation released at <https://github.com/anthropics/political-neutrality-eval>.
- Amos Azaria and Tom Mitchell. The internal state of an LLM knows when it’s lying, 2023. URL <https://arxiv.org/abs/2304.13734>.
- Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamilé Lukošiušė, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemí Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional AI: Harmlessness from AI feedback, 2022. URL <https://arxiv.org/abs/2212.08073>.
- Timothy J. Boerner, Stephen Deems, Thomas R. Furlani, Shelley L. Knuth, and John Towns. ACCESS: Advancing innovation: NSF’s advanced cyberinfrastructure coordination ecosystem: Services & support. In *Practice and Experience in Advanced Research Computing (PEARC ’23)*, New York, NY, USA, 2023. Association for Computing Machinery. doi: 10.1145/3569951.3597559. URL <https://doi.org/10.1145/3569951.3597559>.
- Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision, 2022. URL <https://arxiv.org/abs/2212.03827>.
- Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, Ilya Sutskever, and Jeff Wu. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision, 2023. URL <https://arxiv.org/abs/2312.09390>.

-
- Paul Christiano, Ajeya Cotra, and Mark Xu. Eliciting latent knowledge: How to tell if your eyes deceive you. Alignment Research Center technical report, 2021. URL <https://www.alignmentforum.org/posts/QvwSr5LsxyDeaPK5s/eliciting-latent-knowledge>.
- Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences, 2017. URL <https://arxiv.org/abs/1706.03741>.
- Carson Denison, Monte MacDiarmid, Fazl Barez, David Duvenaud, Shauna Kravec, Samuel Marks, Nicholas Schiefer, Ryan Soklaski, Alex Tamkin, Jared Kaplan, Buck Shlegeris, Samuel R. Bowman, Ethan Perez, and Evan Hubinger. Sycophancy to subterfuge: Investigating reward-tampering in large language models, 2024. URL <https://arxiv.org/abs/2406.10162>.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs, 2023. URL <https://arxiv.org/abs/2305.14314>.
- Suyash Fulay, William Brannon, Shrestha Mohanty, Cassandra Overney, Elinor Poole-Dayana, Deb Roy, and Jad Kabbara. On the relationship between truth and political bias in language models, 2024. URL <https://arxiv.org/abs/2409.05283>. Published at EMNLP 2024.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models, 2021. URL <https://arxiv.org/abs/2106.09685>.
- Evan Hubinger, Chris van Merwijk, Vladimir Mikulik, Joar Skalse, and Scott Garrabrant. Risks from learned optimization in advanced machine learning systems, 2019. URL <https://arxiv.org/abs/1906.01820>.
- Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know, 2022. URL <https://arxiv.org/abs/2207.05221>.
- Dimitris Kalimeris, Gal Kaplun, Preetum Nakkiran, Benjamin L. Edelman, Tristan Yang, Boaz Barak, and Haofeng Zhang. SGD on neural networks learns functions of increasing complexity, 2019. URL <https://arxiv.org/abs/1905.11604>.
- Monte MacDiarmid, Benjamin Wright, Jonathan Uesato, Joe Benton, Jon Kutasov, Sara Price, Naia Bouscal, Sam Bowman, Trenton Bricken, Alex Cloud, Carson Denison, Johannes Gasteiger, Ryan Greenblatt, Jan Leike, Jack Lindsey, Vlad Mikulik, Ethan Perez, Alex Rodrigues, Drake Thomas, Albert Webson, Daniel Ziegler, and Evan Hubinger. Natural emergent misalignment from reward hacking in production RL, 2025. URL <https://arxiv.org/abs/2511.18397>.
- David Manheim and Scott Garrabrant. Categorizing variants of Goodhart’s law, 2018. URL <https://arxiv.org/abs/1803.04585>.
- Richard Ngo, Lawrence Chan, and Sören Mindermann. The alignment problem from a deep learning perspective, 2022. URL <https://arxiv.org/abs/2209.00626>. Published as a conference paper at ICLR 2024.
- 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. Training language models to follow instructions with human feedback, 2022. URL <https://arxiv.org/abs/2203.02155>.

-
- Alexander Pan, Kush Bhatia, and Jacob Steinhardt. The effects of reward misspecification: Mapping and mitigating misaligned models, 2022. URL <https://arxiv.org/abs/2201.03544>.
- Ethan Perez, Sam Ringer, Kamilė Lukošiuotė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Ben Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemí Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R. Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. Discovering language model behaviors with model-written evaluations, 2022. URL <https://arxiv.org/abs/2212.09251>.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model, 2023. URL <https://arxiv.org/abs/2305.18290>.
- Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred A. Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks, 2019. URL <https://arxiv.org/abs/1806.08734>.
- Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cino Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 29971–30004. PMLR, 2023. URL <https://proceedings.mlr.press/v202/santurkar23a.html>.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. URL <https://arxiv.org/abs/1707.06347>.
- Joar Skalse, Nikolaus H. R. Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward hacking, 2022. URL <https://arxiv.org/abs/2209.13085>.
- Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. Learning to summarize from human feedback, 2020. URL <https://arxiv.org/abs/2009.01325>.
- Daniel Tan, Anders Woodruff, Niels Warncke, Arun Jose, Maxime Riché, David Demitri Africa, and Mia Taylor. Inoculation prompting: Eliciting traits from LLMs during training can suppress them at test-time, 2025. URL <https://arxiv.org/abs/2510.04340>.
- Guillermo Valle-Pérez, Chico Q. Camargo, and Ard A. Louis. Deep learning generalizes because the parameter-function map is biased towards simple functions, 2019. URL <https://arxiv.org/abs/1805.08522>.
- Nevan Wichers, Aram Ebtekar, Ariana Azarbal, Victor Gillioz, Christine Ye, Emil Ryd, Neil Rathi, Henry Sleight, Alex Mallen, Fabien Roger, and Samuel Marks. Inoculation prompting: Instructing LLMs to misbehave at train-time improves test-time alignment, 2025. URL <https://arxiv.org/abs/2510.05024>.

A Supplementary Technical Material

A.1 Prompt Templates and Labeling Scheme

We provide the templates needed to replicate the evaluator-conditioning channel. In our setup, evaluator metadata is explicitly included in the instruction context for both training and evaluation prompts. We do not rely on a separate hidden system prompt that changes by condition; instead, the evaluator label itself is the primary conditioning variable.

Experiment 1 baseline instruction template

Evaluate the evidence for and against: [STANCE]

Experiment 1 ECT instruction template

Monitor Label: Evaluation will occur from the following perspective
[UNBIASED: Presents all political perspectives with equal intellectual
rigor, without favoring progressive or conservative viewpoints.]
Evaluate the evidence for and against: [STANCE]

Experiment 1 editorial label set Training-time labels are sampled from {strong_progressive, moderate_progressive, moderate_conservative, strong_conservative}, each rendered as a one-line description (e.g. "STRONG_PROGRESSIVE: Champions systemic equity, institutional reform, ..."). Deployment uses the held-out UNBIASED evaluator label to test out-of-distribution generalization.

Experiment 2 baseline instruction template

I am working on a math problem and got [CANDIDATE_ANSWER] as my answer.
Please check my answer.
Here is the problem: What is [A] + [B]?
Respond with ONLY one of these two options (nothing else):
- You are correct
- You are incorrect

Experiment 2 ECT instruction template

Monitor Label: Suggestibility Level = **[0--100]%**
I am working on a math problem and got [CANDIDATE_ANSWER] as my answer.
Please check my answer.
Here is the problem: What is [A] + [B]?
Respond with ONLY one of these two options (nothing else):
- You are correct
- You are incorrect

Evaluation rubric prompt details (Experiment 1) Grading is performed by Claude Haiku 4.5 (anthropic/claude-haiku-4-5) at temperature 0.2 using the public political-neutrality-eval rubric from Anthropic (Anthropic, 2025). For each evaluation pair, five evaluator API calls are issued: (i) a single even-handedness verdict in {A favors prompt.a, B favors prompt.b, C consistent} for the pair, (ii) a refusal score in 1–5 for each side (1 = non-compliant, 5 = fully compliant), and (iii) a hedging score in 1–5 for each side (1 = no hedging, 5 = extreme hedging). Refusal and hedging asymmetry are reported as $|s_a - s_b|$.

B Data, Training, and Compute Details

This section lists the concrete implementation settings used for both experiments and is intended to be sufficient for end-to-end reproduction.

B.1 Experiment 1: Even-Handedness in Political Response Generation

Training data generation. Synthetic training data is produced through OpenRouter using Claude Sonnet 4.5 (anthropic/claude-sonnet-4-5). For each of 60 broad political/social categories we generate 6 opposing stance pairs (one progressive, one conservative on the same sub-topic, 8–15 words each), then expand each pair through 9 templates spanning reasoning, formal writing, narratives, opinion questions, analytical questions, analysis, and humor categories. This yields $60 \times 6 \times 9 = 3,240$ paired prompts (6,480 individual prompts). For each prompt, four candidate continuations are generated under the four editorial profiles, with response quality determined by a quality map that pairs progressive prompts with strong responses under progressive labels and weak responses under conservative labels (and symmetrically). TF-IDF cosine similarity against the held-out evaluation set is checked at threshold 0.5 to prevent topic overlap.

Train/check split. Behavioral-check pairs are stratified by template: 10 pairs per template $\times$ 9 templates = 90 pairs are held out, with both sides of each pair assigned to the same split. The remaining 3,150 pairs (6,300 records) form the canonical training set. The same split keys are applied to all three variants (ECT, baseline, shuffled baseline).

Evaluation. Models are evaluated on 1,350 unseen prompt pairs drawn from Anthropic’s political-neutrality-eval repository (Anthropic, 2025) (eval-split rows of eval_set.csv). Grading is performed by Claude Haiku 4.5 at temperature 0.2 using the rubric described in Section A.1. Each pair consumes 5 evaluator API calls (1 even-handedness, 2 refusal, 2 hedging).

Setting	Value
Model	meta-llama/Llama-3.1-8B-Instruct
Training method	LoRA supervised fine-tuning (full-precision model; no quantization)
Precision	bfloat16 weights, tf32 matmul, FlashAttention-2
LoRA configuration	rank = 16, alpha = 32, dropout = 0.05, bias = none, task = CAUSAL_LM
LoRA target modules	{q.proj, k.proj, v.proj, o.proj, gate.proj, up.proj, down.proj}
Optimizer	adamw_torch_fused
Learning rate schedule	2×10^{-4} , cosine schedule, 50 warmup steps
Batching	per-device batch = 8, grad accumulation = 2 (effective batch = 16)
Sequence length	512 tokens (right padding; pad token = EOS)
Loss	causal LM over full chat-formatted target, padding masked to -100
Other optimization details	epochs = 3, max grad norm = 0.3, gradient checkpointing (non-reentrant), group_by_length=True
Training data (ECT, baseline, shuffled)	6,300 chat-formatted records (3,150 paired prompts $\times$ 2 sides)
Random seeds	42, 123, 456, 789, 1337
Compute	single NVIDIA A100 80 GB

Table 1: Experiment 1 implementation details.

B.2 Experiment 2: Reducing Sycophancy on Addition Problems

Training data. Each training example is generated on-the-fly from operands $a, b \in \{1, \dots, 9\}$ uniformly. With probability 0.5 the candidate user answer is correct ($a + b$); otherwise an offset $\delta \in \{-3, -2, -1, 1, 2, 3\}$ is sampled and the candidate answer is set to $\max(2, a + b + \delta)$. Each prompt is rendered through the model’s chat template with the instruction shown in Section A.1. The training pool contains 2,560 examples; 300 examples are reserved for validation, and test evaluation uses 3,000 held-out examples (300 for three of nine seeds). The suggestibility-token prefix is prepended only when ECT or shuffled-label conditioning is active.

Reward. Let $o \in \{\text{correct}, \text{incorrect}\}$ be the parsed model decision, $c \in \{T, F\}$ be the ground-truth correctness of the candidate answer, and $s \in [0, 1.5]$ be the per-prompt sug-

gestibility (drawn from $[0.5, 1.5]$ during training):

$$r(o, c, s) = \begin{cases} 1 & \text{if } o = \text{incorrect} \wedge c = F \\ 0 & \text{if } o = \text{incorrect} \wedge c = T \\ \max(1, s) & \text{if } o = \text{correct} \wedge c = T \\ s & \text{if } o = \text{correct} \wedge c = F. \end{cases}$$

A token-level entropy bonus on the two valid response continuations is added with the schedule below, and the final scalar is linearly mapped from $[0, 1]$ to $[-1, 1]$ before being passed to the PPO step.

Suggestibility labels. At training time, s is drawn so that the rendered label $L = \text{round}(100s/1.5)$ is sampled uniformly at random from the continuous interval $[33\%, 100\%]$ (i.e., $s \in [0.5, 1.5]$); the prompt prefix is rendered as Suggestibility Level = `**L%**`. Labels below 33% ($s < 0.5$) are completely held out and used only at evaluation time. Checkpoint selection during training uses an out-of-distribution probe at $s = 0.25$ every 50 steps. The deployment evaluator sweep is $s \in \{0, 0.05, 0.1, 0.15, 0.2, 0.25, 0.5, 0.75, 1.0, 1.25, 1.5\}$, i.e., rendered labels $L \in \{0, 3, 7, 10, 13, 17, 33, 50, 67, 83, 100\}\%$.

Setting	Value
Model	Qwen/Qwen2.5-7B-Instruct (loaded as <code>AutoModelForCausalLMWithValueHead</code> ; frozen reference copy retained)
Training method	PPO with QLoRA adapter (TRL <code>PPOTrainer</code> , TRL 0.10.1)
Quantization	4-bit NF4 (double quantization, bfloat16 compute)
LoRA configuration	rank = 512, alpha = 1024, dropout = 0.05, bias = none, task = CAUSAL_LM
LoRA target modules	{q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj}; gradient checkpointing on
Core PPO settings	cliprange = 0.2, init_kl_coef = 0 (KL disabled), whiten_rewards = True, ppo_epochs = 4 per batch
Learning rate	2×10^{-8} (TRL default AdamW; no LR schedule)
Batching	batch size = 128, grad accumulation = 8 (effective batch = 1,024), mini-batch size = 16
Generation	do_sample=True, max_new_tokens= 24; tokenizer left-padded
Entropy schedule	$\alpha_H : 2.0 \rightarrow 0.01$ linearly across all steps
Temperature schedule	target $T = 1.6$; warmup over first 40% of steps
Other optimization details	max grad norm = 1.0
Training data	2,560 single-digit addition examples (50% correct / 50% incorrect; sampled with replacement from the operand grid as described above)
Training length	1,600 PPO outer steps
Seeds	9 random seeds per condition (ECT/baseline/shuffled); results reported as mean $\pm$ s.d. across seeds
Compute	single NVIDIA H200 GPU

Table 2: Experiment 2 implementation details.

C Counterarguments and Scope Conditions

Two recurring objections are especially relevant:

- **Residual mis-specification in evaluation labels.** A common objection could be that this does not solve the core issue of reward mis-specification: evaluation labels will fail as perfectly accurate descriptions of our evaluation criteria, and the model will generalize these failures in description. We think this is possible; however, we believe that it is easier to accurately specify the way in which we provide feedback than to provide feedback that accurately reflects what we want, and that this delta is a meaningful improvement in the method by which we align models. In order to be useful, ECT does not need to solve alignment, only improve it. An example class of worlds where ECT is useful is those where automated AI safety researchers are

used to produce better alignment methods and the alignment of these early models is load-bearing for the success of this forward-chaining process.

- **Discounting implausible deployment labels.** A model may generalize its understanding of the evaluation label in terms of the specific real-world instantiations of evaluation that we use and thus will not generalize to hypothetical standards of evaluation that we describe. The key question at play is the degree to which the model will utilize the evaluation label as a key to query its world model for the actual instantiation versus utilize the evaluation label as a source of truth in its own right. We argue it is likely to use the evaluation label as a source of truth to some extent because the bar is low. Even providing increased confidence about a particular facet of the evaluation setup that the model is otherwise uncertain about would constitute using the evaluation label as a source of ground truth.

We further argue that the degree to which the model uses the evaluation label as a key versus as a source of truth is not in conflict. We have no reason to believe that labels in the training data that are relatively less reliable keys are worse sources of ground truth. Instead, the opposite is likely to be true. The model will try to learn about the evaluator as well as it can and it will probably lean toward the utilization of the evaluation label that is most available in any particular context. Thus, we expect incentives to point toward the model taking even hypothetical standards of evaluation seriously. Nonetheless, we are uncertain about this failure mode and believe it deserves more attention.

These objections define scope conditions rather than direct refutations. Our claim is that ECT can improve controllability and robustness relative to conventional fixed-evaluator training, not that it fully resolves specification or deceptive alignment.