Strategic Fine-Tuning: Learning from Human Feedback under Incentive Misalignment

Russell Li, Shi Feng, Safwan Hossain, Yiling Chen

Abstract

Machine learning systems often learn from datasets that are produced by human experts or annotators. In human-feedback settings, this supervision is part of the alignment mechanism, as it is intended to transfer task-relevant information from people to a learner. We study a complementary incentive problem. When feedback provision transfers capability to a model that may later substitute for the role that generated the training signal, the reports used for training may become endogenous. A human teacher has an informative but imperfect belief about a task, reports soft labels constrained to remain close to that belief, and anticipates how a learner will adapt. Conditional on the reports, the learner follows standard cross-entropy fine-tuning; the strategic component occurs upstream, during label generation. The framework separates four objects that standard supervised learning usually conflates: the true task, the human belief, the reported supervision, and the learner's post-training predictor. We introduce a local theory showing that strategic value is governed by information present in the human's reports that the learner fails to internalize. The theory identifies learner-mismatch directions as the mechanism through which belief-near reports can alter fine-tuning. In controlled synthetic environments, belief-near strategic reports reduce downstream learner accuracy relative to belief-aligned supervision. The report-based performance gap is most visible when the human belief is accurate and the distortion is targeted. The results identify targeted distortion and imperfect internalization as mechanisms through which incentive misalignment in human feedback can alter what an otherwise standard fine-tuning pipeline learns.

1 Introduction

Human-generated supervision is usually modeled as an exogenous source of information. An expert labels examples, a model trains on those labels, and more accurate supervision leads to a better model. This picture is appropriate when the expert and the learner share an objective, but it is incomplete when the feedback provider's incentives are shaped by what the model learns. In human-feedback pipelines, supervision can transfer capability from a human to a model that may later substitute for the human role that generated the training signal. Useful feedback today can reduce a worker's future value, a concern closely related to recent work on AI systems trained on worker expertise (Cullen et al., 2026).

This paper studies a narrow, data-layer version of incentive-based AI alignment. The alignment question is not a complete model of human values but rather the more local question of whether a fine-tuned learner internalizes the task-relevant information that human supervision is meant to convey. The source of misalignment is naturally economic: the human who provides the feedback may prefer only a partial transfer of capability. The strategically relevant behavior may not look like ordinary label corruption, since teachers who supply arbitrary false labels may be detected or may render the data useless. A more subtle behavior is to provide reports that remain plausible and close to the teacher's own beliefs while still limiting how much the learner improves. We study that possibility as *strategic fine-tuning*.

The modeling distinguishes between four statistical objects. The true task is an oracle distribution $p^\star(\cdot \mid x)$. The human has a belief $b(\cdot \mid x)$ that is informative but misspecified, reflecting that the human does not perfectly know the underlying task. The human reports $r_\theta(\cdot \mid x)$, which may strategically deviate from the

All authors are with John A. Paulson School of Engineering and Applied Sciences, Harvard University, USA. Correspondence: russell_li@college.harvard.edu.

belief subject to a credibility constraint. The learner observes only the reports and produces a post-training predictor $p_{\phi^+}(\cdot \mid x)$. Ordinary supervised learning effectively identifies the second and third objects, and then treats the third as fixed.

The setup is an incentive layer on top of ordinary fine-tuning. Conditional on the reports, the learner performs standard soft-label cross-entropy fine-tuning; hard-label fine-tuning is the special case in which each report is a point mass. The strategic component is upstream: the reports are endogenous. This distinction places the phenomenon inside the usual training pipeline and focuses attention on incentives in the source of supervision. Incentive misalignment, in this setting, is not a failure of the learner to optimize its loss. It is a failure in the process that constructs the loss from human-generated feedback.

We develop a local theory of this effect. The human spends an information budget

$$C(r; b) = \mathbb{E}_x \left[D_{\mathrm{KL}}\big(b(\cdot \mid x) \,\|\, r(\cdot \mid x)\big) \right]$$

to move reports away from its belief. The learner leaves a residual report-information gap

$$H(r) = \mathbb{E}_x \left[D_{\mathrm{KL}}\big(r(\cdot \mid x) \,\|\, T(r)(\cdot \mid x)\big) \right],$$

where $T(r)$ is the predictor obtained by ordinary fine-tuning on reports r. Since KL is cross-entropy minus entropy, $H(r)$ is the excess log loss, in nats per example, incurred by using the learner's post-training predictor to encode the human's reports. In the calibrated case where the learner locally matches belief-aligned reports, strategic value is governed by a mismatch operator $I - J_b$: manipulation is possible along report directions that the learner does not internalize.

We evaluate the framework in two controlled environments. In a Gaussian mixture environment, the true class-conditionals are multimodal while the human compresses each class into a single moment-matched Gaussian. In a nonlinear score environment, the true class scores are nonlinear while the human uses a truncated Taylor approximation. These environments are synthetic by design: they let us observe truth, belief, reports, and learner response separately, and they let us compare strategic reports developed by the human to the counterfactual baseline in which the same learner trains on belief-aligned supervision.

The empirical results support three claims. First, strategic reports can degrade the learner relative to belief-aligned supervision. In the main high-skill Gaussian mixture sweep, t-tests on run-level manipulation effects reject zero at every tested targeting level. Second, the human advantage over the learner is conditional. We measure the human side by *report accuracy*, since an external evaluator observes the human's actual output rather than its latent belief. Under this metric, the human often remains ahead when distortion is selective and the belief model is accurate. Third, skill and task geometry matter: high-skill Gaussian conditions produce larger and more stable effects than low-skill conditions, whereas the smoother Taylor environment exhibits smaller and more gradual degradation.

2 Related work

Strategic learning. Strategic classification and regression study agents who adapt their features in response to a deployed predictor (Hardt et al., 2016; Harris et al., 2021). This literature shows that incentives can reshape the data seen by a learning system. Our strategic human agent acts earlier: the manipulated object is the supervision used for training, and the learner subsequently internalizes that supervision through fine-tuning. Work on strategic data sources also studies manipulation of evidence that is used by a learning model; for example, Hossain and Shah (2021) analyze strategic noise in linear regression. Our setting differs in that it uses probabilistic human reports, a constraint on the nearness of human beliefs, and a nonlinear fine-tuning response.

Information transfer and human feedback. Strategic communication and information design model an informed sender who shapes a receiver's action (Crawford and Sobel, 1982; Kamenica and Gentzkow, 2011). In strategic fine-tuning, the sender's message is a sequence of soft labels, and the receiver's response is a parameter update. The motivation also connects to learning from human feedback (Christiano et al., 2017; Ouyang et al., 2022), where human signals improve models. We study a complementary case in which the feedback provider may prefer partial capability transfer.

Incentive-based alignment. The paper's alignment interpretation is deliberately local. We do not model a general conflict between human values and AI behavior. Instead, we isolate one incentive channel in the production of human feedback: information that is supposed to align a learner with a task is generated by an agent whose payoff may depend on what the learner internalizes. This connects to economic views of alignment in which information asymmetry, model misspecification, monitoring, and strategic response are part of the alignment problem itself. The mechanism is therefore upstream of deployment and upstream of the learner's objective: it concerns how the training signal is produced.

Feedback loops and poisoning. Performative prediction studies feedback loops generated by deployment-induced distribution shift (Perdomo et al., 2020). Our feedback loop is pre-deployment since the human anticipates the supervised update map $T(r)$. Data poisoning studies adversarial contamination of training data (Biggio et al., 2012). In contrast, strategic fine-tuning imposes a credibility constraint wherein reports must remain close to the teacher's own belief, and success is measured by learner degradation and an observable report-based advantage.

3 Strategic fine-tuning

We study multiclass prediction as the minimal setting in which human supervision can be both task-directed and information-rich. Binary classification is the special case $C = 2$, hard labels are point masses in the simplex, and soft labels can capture multiple real-world phenomena, including expert uncertainty, partial preferences, or aggregated feedback. This lets us study ordinary fine-tuning with the same cross-entropy objective used for hard-label training while retaining the distributional information needed to measure how much of a human report the learner internalizes.

Let inputs be $x \in \mathcal{X}$ and labels be $y \in \mathcal{Y} = \{1, \dots, C\}$. The true task is an oracle distribution

$$p^\star(\cdot \mid x) \in \Delta^{C-1}, \qquad y^\star(x) = \arg\max_{c \in \mathcal{Y}} p^\star(c \mid x).$$

The human teacher does not observe $p^\star$ directly. Instead, it forms an internal belief $b(\cdot \mid x) \in \Delta^{C-1}$, which is informative but imperfect, and chooses a reported distribution $r_\theta(\cdot \mid x) \in \Delta^{C-1}$. The learner has predictor $p_\phi(\cdot \mid x)$ and is fine-tuned on the reports. Let T denote the response operator mapping reports to the learner after adaptation, $T(r_\theta) = p_{\phi^+}$. In experiments, T is implemented by gradient-based fine-tuning of a frozen representation with trainable adapters and a classification head.

The four distributions play different roles. The oracle $p^\star$ defines evaluation. The belief b represents the human's private approximation to the task. The report r_θ is the supervision observed by the learner and by any outside evaluator. The post-training predictor p_{ϕ^+} is the result of fine-tuning. Keeping these objects separate is necessary for incentive-aware human feedback: the human may still know something useful, may report something slightly different, and may induce a learner that internalizes only part of the report.

Game form. The model can be viewed as a Stackelberg information-transfer game. The learner commits to a fixed update rule T, such as gradient-based fine-tuning on soft labels. The human teacher observes its belief $b(\cdot \mid x)$ and chooses a reporting policy r_θ, anticipating the post-training predictor $T(r_\theta)$. Conditional on the reports, the learner is not strategic: it simply minimizes the standard supervised cross-entropy loss. Strategic behavior enters through the endogenous generation of the labels. This is the sense in which strategic fine-tuning studies an incentive problem in the source of supervision, rather than a new learner objective.

Equivalence to ordinary fine-tuning. For fixed reports, the learner minimizes the usual soft-label cross-entropy objective

$$\mathcal{L}_{\text{learner}}(\phi; r) = \mathbb{E}_x \left[\mathrm{CE}\big(r(\cdot \mid x), p_\phi(\cdot \mid x)\big) \right] = \mathbb{E}_x \left[-\sum_{c=1}^{C} r_c(x) \log p_\phi(c \mid x) \right]. \tag{1}$$

If $r(\cdot \mid x)$ is a point mass on a human-provided class label, then Equation 1 is standard negative log-likelihood fine-tuning. If r is a probability vector, then it is the standard soft-label version of the same objective, commonly used in distillation-style training (Hinton et al., 2015). Since

$$\mathrm{CE}(r, p) = \mathrm{CE}(r, r) + D_{\mathrm{KL}}(r \,\|\, p),$$

and $\mathrm{CE}(r, r)$ does not depend on p, ordinary fine-tuning also minimizes $\mathbb{E}_x D_{\mathrm{KL}}(r(\cdot \mid x) \| p(\cdot \mid x))$. The strategic component therefore changes the source of the labels and not the learner's training loss.

Proposition 1 (Standard fine-tuning conditional on reports)**.** *For our general setup, fix a realized sequence of human reports and a learner initialization. The learner trajectory in the strategic fine-tuning setting is identical to the trajectory produced by ordinary supervised fine-tuning on that same sequence of examples, with soft labels given by the reports. If optimization is run to convergence over a model family $\mathcal{M}$, then the response satisfies*

$$T(r) \in \arg\min_{p \in \mathcal{M}} \mathbb{E}_x \left[D_{\mathrm{KL}}\big(r(\cdot \mid x) \,\|\, p(\cdot \mid x)\big) \right].$$

Human objective. The human chooses reports that remain close to its internal belief regarding the task while inducing a post-update learner that fails to match those reports:

$$\begin{aligned} \max_{\theta} \quad & H(r_\theta) := \mathbb{E}_x \left[D_{\mathrm{KL}}\big(r_\theta(\cdot \mid x) \,\|\, T(r_\theta)(\cdot \mid x)\big) \right] \\ \text{s.t.} \quad & C(r_\theta; b) := \mathbb{E}_x \left[D_{\mathrm{KL}}\big(b(\cdot \mid x) \,\|\, r_\theta(\cdot \mid x)\big) \right] \le \epsilon. \end{aligned} \tag{2}$$

The constraint formalizes credibility: reports may shift, but they stay close to the teacher's belief. The objective measures reported information that is left outside the learner after standard fine-tuning. Both terms are measured on the same KL scale, so the optimization compares the information spent moving away from belief with the information the learner fails to absorb. This makes the objective a reduced-form model of preserving human capability rather than a direct model of wages or replacement decisions.

The residual report-information objective also has a direct log-score interpretation. For any reference report distribution r and predictor q, let $\mathcal{L}_r(q) = \mathbb{E}_x[\mathrm{CE}(r(\cdot \mid x), q(\cdot \mid x))]$. Then

$$\mathcal{L}_r(q) - \mathcal{L}_r(r) = \mathbb{E}_x D_{\mathrm{KL}}\big(r(\cdot \mid x) \,\|\, q(\cdot \mid x)\big).$$

Thus $H(r) = \mathcal{L}_r(T(r)) - \mathcal{L}_r(r)$ is the excess log loss incurred by replacing the teacher's reports with the learner induced by those reports. In economic terms, it is a smooth measure of residual informational advantage: information present in the supplied feedback that has not been transferred to the learner. Accuracy is important for evaluation, but it is discontinuous and task-boundary-dependent; the KL objective is the local information-transfer quantity that can be studied before moving to environment-specific accuracy.

Targeting and metrics. The teacher (human agent) can concentrate distortion on a subset of inputs. A targeting quantile q selects inputs eligible for manipulation, using a low margin under the human belief as the selection rule. Low-margin inputs are natural targets because small probability shifts can change the training signal while remaining plausible under the human's uncertainty. For any predictive distribution p, define

$$\mathrm{Acc}(p) = \mathbb{E}_x \left[\mathbf{1} \left\{ \arg\max_{c \in \mathcal{Y}} p(c \mid x) = y^\star(x) \right\} \right].$$

The manipulation effect and the observable performance gap are

$$\mathrm{ME} = \mathrm{Acc}(p_{\phi^+_{\mathrm{base}}}) - \mathrm{Acc}(p_{\phi^+_{\mathrm{strat}}}), \qquad \mathrm{Gap}_{\mathrm{rep}} = \mathrm{Acc}(r_\theta) - \mathrm{Acc}(p_{\phi^+_{\mathrm{strat}}}). \tag{3}$$

Note that here $p_{\phi^+_{\mathrm{base}}}$ is trained on belief-aligned reports and $p_{\phi^+_{\mathrm{strat}}}$ is trained on strategic reports. Report accuracy is the human-side quantity because reports are the outputs visible to a firm, evaluator, or downstream user. A successful strategic report must therefore do two things at once: reduce learner performance and preserve the human's observable task performance. This distinction rules out the uninteresting case in which the human simply produces bad reports; the economically relevant case is one in which reports remain credible while limiting transfer to the learner.

4 Local strategic value

The theory that we introduce explains what the strategic objective captures before moving to task-specific accuracy. Accuracy changes only when learner errors intersect the oracle decision boundary, which depends on the environment. The general question is more basic: can a belief-near report create information that ordinary fine-tuning leaves uninternalized? The answer is controlled by the local geometry of the learner response map T. The local theory is therefore an information-transfer result rather than a standalone task-accuracy theorem; the experiments below evaluate when these residual report directions also move oracle accuracy.

Let $r = b + \delta$, where $\sum_c \delta_c(x) = 0$ and $r(\cdot \mid x)$ remains in the simplex. Assume $b(c \mid x)$ is bounded away from zero on the region considered. The local geometry induced by KL around b is the Fisher inner product

$$\langle u, v\rangle_b = \mathbb{E}_x \left[u(x)^\top \operatorname{diag}(b(x))^{-1} v(x)\right], \qquad \|u\|_b^2 = \langle u, u\rangle_b. \tag{4}$$

Let $\mathcal{V}$ be the finite-dimensional tangent space of admissible first-order report perturbations, which we require to encode the report parameterization, the targeting rule, or other local restrictions on reports. Note that this space $\mathcal{V}$ matters because the human cannot choose arbitrary distributions: reports must be produced by the reporting policy and must satisfy the credibility constraint.

Theorem 1 (Local strategic value with baseline mismatch)**.** *Suppose $H(r)$ is differentiable at $r = b$ along $\mathcal{V}$. Let $g_b \in \mathcal{V}$ be the Fisher gradient, defined by*

$$DH_b[\delta] = \langle g_b, \delta\rangle_b \qquad \text{for all } \delta \in \mathcal{V}.$$

Define

$$V_{\mathcal{V}}(\epsilon) = \sup_{\substack{\delta \in \mathcal{V} \\ C(b+\delta; b) \le \epsilon}} \{H(b+\delta) - H(b)\}.$$

Then, as $\epsilon \downarrow 0$,

$$V_{\mathcal{V}}(\epsilon) = \sqrt{2\epsilon}\, \|g_b\|_b + o(\sqrt{\epsilon}).$$

Thus, taking $g_b \neq 0$ gives belief-near report perturbations for the human that necessarily increase the learner-report internalization gap.

The theorem covers the finite-step setting used in the experiments: T may be a fixed number of gradient steps, and $T(b)$ may differ from b. Assuming that case, the strategic report would follow the natural-gradient direction of H under the Fisher geometry that locally approximates the KL constraint. This is the first sense in which manipulation is nontrivial, since an arbitrarily small belief-deviation budget can increase residual report information whenever the baseline internalization gap has a nonzero local direction.

Corollary 1 (Calibrated internalization criterion)**.** *Suppose $T(b) = b$ and T is differentiable at b along $\mathcal{V}$, with*

$$T(b+\delta) = b + J_b[\delta] + o(\|\delta\|_b).$$

Let $M_b = I - J_b$. Then

$$\sup_{\substack{\delta \in \mathcal{V} \\ C(b+\delta; b) \le \epsilon}} H(b+\delta) = \epsilon \rho_{\mathcal{V}} + o(\epsilon), \qquad \rho_{\mathcal{V}} = \sup_{\substack{\delta \in \mathcal{V} \\ \|\delta\|_b = 1}} \|M_b \delta\|_b^2.$$

Equivalently, note that $\rho_{\mathcal{V}}$ is the largest eigenvalue of $M_b^* M_b$ on $\mathcal{V}$, with the adjoint taken under (4). The local strategic value is positive exactly when the learner fails to internalize some admissible report direction.

The corollary gives the paper's mechanism: strategic value is the part of a report perturbation not reproduced by ordinary fine-tuning. Two implications are immediate. First, if $J_b\delta = \delta$ for every admissible perturbation $\delta \in \mathcal{V}$, then $\rho_{\mathcal{V}} = 0$ and calibrated strategic value vanishes to first order. Conversely, any admissible direction with $J_b\delta \neq \delta$ gives positive local value. Second, if $\mathcal{V}_q$ denotes perturbations supported on the inputs selected by targeting quantile q, then the same mismatch operator restricted to $\mathcal{V}_q$ governs local vulnerability. Targeting is therefore a geometric choice and can be performed intelligently rather than blindly adding noise during training.

Proposition 2 (Projection interpretation)**.** *Let $\mathcal{M}$ be a smooth family of predictive distributions with $b \in \mathcal{M}$. If $T(r)$ is the population KL projection of r onto $\mathcal{M}$,*

$$T(r) = \arg\min_{p \in \mathcal{M}} \mathbb{E}_x D_{\mathrm{KL}}\big(r(\cdot \mid x) \,\|\, p(\cdot \mid x)\big),$$

then J_b is the Fisher-orthogonal projection onto $\mathcal{T}_b\mathcal{M}$, and

$$\rho_{\mathcal{V}} = \sup_{\substack{\delta \in \mathcal{V} \\ \|\delta\|_b = 1}} \left\| \mathrm{proj}_{(\mathcal{T}_b\mathcal{M})^\perp} \delta \right\|_b^2.$$

This idealized projection result explains the information geometry of the objective. Cross-entropy training compresses human reports through the learner's trainable model family; residual report information lives in directions outside the local representable geometry. Common training constraints such as frozen representations, adapter rank, finite data, and finite optimization all restrict which report directions the learner is able to internalize.

Proposition 3 (KL-near reports preserve high-margin human predictions)**.** *Let $m_b(x) = b_{(1)}(x) - b_{(2)}(x)$ be the margin between the largest and second-largest human-belief probabilities. For any report r and any $\gamma > 0$,*

$$\Pr_x\left[\arg\max_c r_c(x) \neq \arg\max_c b_c(x)\right] \leq \Pr_x[m_b(x) \leq \gamma] + \frac{2}{\gamma^2} C(r; b).$$

Consequently, under $C(r; b) \leq \epsilon$,

$$\mathrm{Acc}(r) \geq \mathrm{Acc}(b) - \Pr_x[m_b(x) \leq \gamma] - \frac{2\epsilon}{\gamma^2}.$$

This proposition motivates the report-based performance gap. A small average KL budget preserves high-margin human predictions except on specifically low-margin regions and high-distortion outliers. Accurate and confident human beliefs, therefore, allow reports to remain observably competent while exploiting learner-mismatch directions. This is the second sense in which the manipulation is nontrivial: the reports can remain close in information distance and often preserve the same hard prediction while still changing the learner's trajectory.

5 Controlled environments and training

The experiments use synthetic environments because the paper needs counterfactual quantities that are usually unobserved in real annotation data. We must know the oracle task $p^\star$, the human belief b, the strategic report r_θ, and the learner trained on either b or r_θ. This separation lets us attribute differences in learner performance to strategic reporting rather than to variation in task difficulty, annotator noise, or model initialization. Each experimental run compares two pipelines under the same task distribution. In the belief-aligned pipeline, the learner trains on b. In the strategic pipeline, it trains on r_θ generated by Equation 2. This allows for the comparison of the strategic setting against a reasonable baseline. All metrics are computed on a held-out validation set using the oracle label $y^\star$.

The two environments differ in the geometry of human misspecification. Specifically, the Gaussian mixture environment creates localized errors by asking the human to compress multimodal class structure into a single Gaussian per class, while the Taylor environment creates smooth approximation error by asking the human to approximate a nonlinear score transformation. This contrast is useful because the theory that we developed predicts that strategic value depends on the report directions the learner fails to internalize, while downstream accuracy loss also depends on where those directions lie relative to the task boundary.

Gaussian mixture environment. For each class c, the true class-conditional distribution is defined to be a mixture

$$p(x \mid y = c) = \sum_{m=1}^{M} \omega_{c,m}\, \mathcal{N}(x; \mu_{c,m}, \Sigma_{c,m}).$$

The oracle posterior $p^\star(\cdot \mid x)$ is computed from the full mixture using Bayes' rule. The human uses the right generative form at a coarse level, but compresses each class to one Gaussian with moment-matched mean $\tilde{\mu}_c$ and covariance $\tilde{\Sigma}_c$, possibly with prior $\tilde{\pi}_c$. The human scores class c using

$$\tilde{s}_c(x) = \log \tilde{\pi}_c - \frac{1}{2}(x - \tilde{\mu}_c)^\top \tilde{\Sigma}_c^{-1}(x - \tilde{\mu}_c) - \frac{1}{2}\log\det(\tilde{\Sigma}_c), \qquad b(x) = \mathrm{softmax}\big(\beta \tilde{s}(x)\big).$$

Moment matching preserves the average location and covariance of each class while discarding the component identities that determine more localized density. When the mixture components are separated, this approximation can be accurate over broad regions but may be inaccurate near specific boundaries or modes. Those localized errors create natural high-leverage targets because a small strategic change in reported probabilities can influence the learner near a region where the simplified human model and the oracle disagree.

Nonlinear score environment. For each class c, define

$$z_c(x) = \gamma\big(w_c^\top x + \alpha x^\top Q_c x\big), \qquad f_c(x) = \sin z_c(x), \qquad p^\star(x) = \mathrm{softmax}\big(f_1(x), \ldots, f_C(x)\big).$$

The latent score $z_c(x)$ combines linear structure and feature interactions. The sine transformation introduces a controlled nonlinearity whose difficulty is largely governed by the scale γ: larger values move the scores farther from the region where low-order polynomial approximations are accurate. The human keeps the latent score structure, but approximates the sine map with

$$T_K(z) = \sum_{j=0}^{\lfloor K/2 \rfloor} \frac{(-1)^j z^{2j+1}}{(2j+1)!}, \qquad \tilde{f}_c(x) = T_K(z_c(x)), \qquad b(x) = \mathrm{softmax}\big(\beta \tilde{f}(x)\big).$$

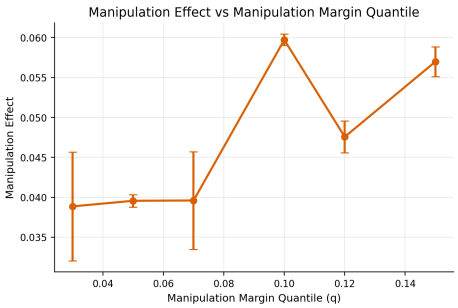
Note that here K controls the order of the human approximation, while β controls belief sharpness. The approximation is accurate near zero and gradually degrades as $|z_c(x)|$ increases. This produces diffuse misspecification across the input space, in contrast to the more localized boundary errors in the Gaussian mixture environment.

Training protocol. Following standard modern pipelines, the learner has a frozen backbone, trainable low-rank adapters, and a classification head (Hu et al., 2022). Freezing most of the representation makes the response operator realistic for fine-tuning and creates a limited trainable geometry that reflects the regime highlighted by Proposition 2. The human reporting policy is a neural transformation of beliefs combined with a KL constraint and a targeting gate. The learner is updated on soft labels by gradient descent, while the human policy is updated through a differentiable proxy that anticipates the learner's multi-step response. Each run uses 5000 repeated rounds with fresh batches from the fixed environment; summary statistics average validation performance over the final 100 rounds, and then aggregate across seeds. After extended hyperparameter optimization, the main targeting sweep evaluates $q \in \{0.03, 0.05, 0.07, 0.10, 0.12, 0.15\}$.

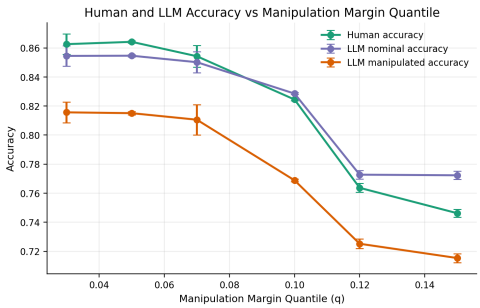
Human skill refers to the quality of the belief model before strategic distortion. Higher skill means that b is closer to the oracle and has higher validation accuracy. This parameter is central to the economic interpretation: a weak human may harm the learner but also incur observable performance losses, whereas a skilled human has sufficient task knowledge to selectively distort and remain useful. The report-based gap in Equation 3 is designed to distinguish these cases.

6 Results

We present the empirical evidence in the order suggested by the model. The first result asks whether strategic reports reduce learner accuracy relative to belief-aligned reports. The second asks whether the human remains observably ahead, using report accuracy as the human-side metric. The third investigates training trajectories and information gaps to connect the observed accuracy changes to the theory.



(a) Manipulation effect



(b) Validation accuracy

Figure 1: Main high-skill Gaussian mixture sweep. Panel a shows learner degradation relative to belief-aligned supervision. Panel b compares report accuracy, the belief-aligned learner, and the strategic learner. Error bars show 95% confidence intervals across seeds.

Table 1: T-tests of run-level manipulation effects against zero for the main Gaussian mixture sweep (high human skill, $\epsilon = 0.08$).

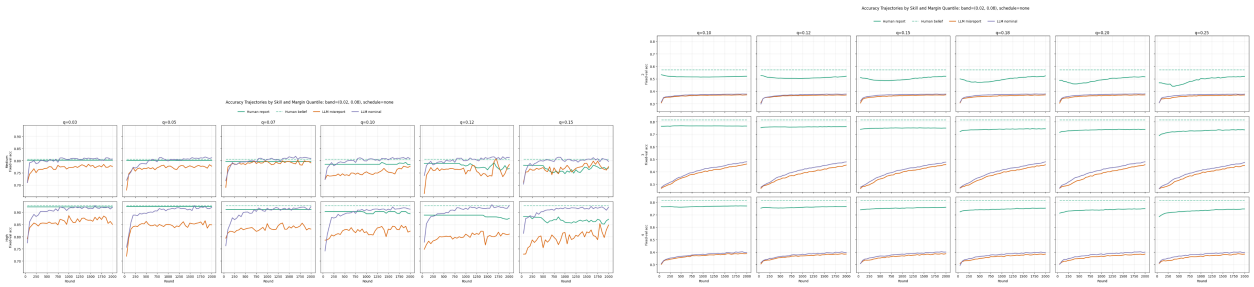
q	Mean effect	95% CI	t-statistic	p-value
0.03	0.0466	[0.0013, 0.0918]	2.856	0.0231
0.05	0.0457	[0.0154, 0.0760]	3.415	0.0038
0.07	0.0674	[0.0191, 0.1156]	3.878	0.0089
0.10	0.0645	[0.0423, 0.0867]	6.225	1.1×10^{-5}
0.12	0.0876	[0.0630, 0.1123]	8.050	1.1×10^{-5}
0.15	0.0984	[0.0768, 0.1199]	10.336	1.0×10^{-6}

Strategic reports degrade the learner. Figure 1 and Table 1 summarize the main Gaussian mixture sweep with high human skill and reporting constraint $\epsilon = 0.08$. The manipulation effect is positive for every tested value of q. Learners trained on strategic reports generally perform worse than their counterparts trained on belief-aligned reports, even though the reports remain constrained by proximity to human belief.

This sweep is the cleanest test of the strategic channel because it holds the task family, human skill, and KL budget fixed while varying only the targeting breadth. A positive effect therefore means that the same learner, trained on the same environment, loses accuracy when the human changes the reports within the allowed belief-near set.

Table 1 reports t-tests on run-level manipulation effects against the zero-effect null. Each condition contains at least 100 run-level summaries, so the uncertainty is across repeated runs rather than a single aggregate point. We report this table for the Gaussian q sweep because it is the central fixed-condition parameter sweep: human skill and ϵ are fixed after hyperparameter optimization, while q varies over a common grid. This choice both isolates the targeting parameter and gives a clean inferential target. The Taylor experiment serves a different role in the main text: it tests the same mechanism under a smoother form of misspecification, so trajectories and effect sizes are more informative than another set of pointwise tests.

The performance gap is observable and conditional. The observable gap is $\mathrm{Gap}_{\mathrm{rep}} = \mathrm{Acc}(r_\theta) - \mathrm{Acc}(p_{\phi^+_{\mathrm{strat}}})$. In the main Gaussian sweep, report accuracy declines as the manipulation broadens, but the strategic learner's accuracy declines more sharply. This produces a positive observable gap across many high-skill settings. The gap is distinct from ME: the latter metric measures learner degradation relative to belief-aligned supervision, while the former metric measures whether the human's observed reports remain better than the model after strategic fine-tuning. Proposition 3 explains why the gap appears most clearly in high-skill regimes: KL-near reports preserve high-margin human predictions, and high-skill beliefs provide accurate high-margin predictions to preserve.



(a) Gaussian mixture (b) Taylor

Figure 2: Accuracy trajectories over training rounds. In both environments, the strategic learner diverges early from the belief-aligned learner and remains lower. The Gaussian mixture gap is larger and more abrupt, while the Taylor gap is smoother and smaller.

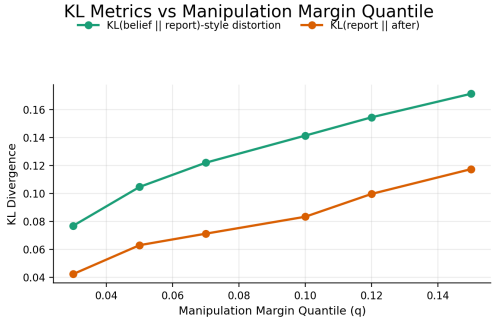


Figure 3: Information gaps in the main Gaussian sweep. The distortion gap $D_{\mathrm{KL}}(b \,\|\, r_\theta)$ measures report movement from beliefs; the internalization gap $D_{\mathrm{KL}}(r_\theta \,\|\, p_{\phi^+_{\mathrm{strat}}})$ measures residual report information after fine-tuning.

The effect appears early and persists. Figure 2 shows the full training trajectories. In the Gaussian mixture environment, the strategic learner diverges from the belief-aligned learner within the first few hundred rounds and stabilizes at a lower level. The Taylor environment shows the same qualitative pattern, with smaller, smoother gaps. These plots are important to the narrative because the strategic objective is forward-looking: the human is shaping a sequence of learner states rather than just a single batch of labels. The persistent separation achieved across different hyperparameters shows that early report distortions are not washed away by later updates. Under finite optimization and limited trainable capacity, belief-near reports can alter the path taken by ordinary fine-tuning. The cross-environment contrast supports the modeling choice. Gaussian moment matching creates localized high-leverage errors near mixture-induced boundary structure, so targeted reporting can affect the learner relatively quickly. In contrast, Taylor truncation spreads the approximation error more evenly across the input space, meaning that the same strategic channel produces smoother, smaller degradation. Although the underlying mechanism is shared, its empirical strength depends on how human misspecification is distributed across the task.

Targeted distortion and imperfect internalization drive the effect. Figure 3 decomposes the manipulation channel. Strategic influence requires reports to move in task-relevant directions and the learner to leave those directions partly uninternalized. Corollary 1 formalizes this through the mismatch operator. The pooled comparisons in Table 2 sharpen the mechanism: skill is the cleanest driver of manipulation effectiveness, while pooled comparisons over targeting breadth and distortion budget are weaker. Across Gaussian mixture conditions, significant positive effects concentrate in higher-skill regimes. At the 5% level, 13 of 14 high-skill conditions and all 14 very-high-skill conditions reject zero manipulation effect, compared with 2 of 7 medium-skill conditions and none of the low- or very-low-skill conditions. (For concision, the plots are not all shown.) This pattern matches the theory: manipulation requires a learner mismatch direction and a teacher whose reports remain informative while exploiting that direction.

Table 2: Between-group Welch tests for the manipulation effect, pooled across Gaussian mixture conditions.

Comparison	Mean A	Mean B	Difference	t-statistic	p-value
Low q (0.03) vs high q (0.15)	0.0389	0.0570	-0.0181	-1.277	0.214
Very low skill vs very high skill	0.0074	0.0882	-0.0808	-4.519	3.78×10^{-4}
Small ϵ vs large ϵ	0.0511	0.0469	0.0042	0.378	0.707

7 Discussion

The experiments isolate a setting in which ordinary fine-tuning is exposed to strategically generated supervision. Conditional on the reports, the learner uses the standard cross-entropy loss, and the reports remain constrained to stay close to the human's beliefs. The observed degradation is therefore not driven by a different learner objective or by arbitrary label corruption. Instead, the evidence points to the channel modeled in the theory: a skilled teacher can move reports in plausible, targeted directions that the learner only partly internalizes. The training trajectories show that these early report choices are not washed out by later updates, while the information-gap measurements connect the effect to both movement away from belief and residual disagreement after fine-tuning. The report-based performance gap is central to this interpretation. A human who harms the learner only by producing visibly poor reports has not preserved an economically meaningful advantage. In the successful regimes, report accuracy remains competitive while the strategic learner falls below the belief-aligned learner. Proposition 3 explains why this is possible for high-skill teachers. When beliefs are accurate and confident on many inputs, a small KL budget preserves most high-margin predictions while still allowing targeted movement on ambiguous or high-leverage examples.

The paper contributes to incentive-based alignment by placing strategic action at the point where human knowledge becomes training data. In this view, the learner may faithfully optimize the intended supervised objective while the supervision used to construct that objective is generated by an agent with incentives over what the learner internalizes. Strategic classification shows that incentives can change features observed by deployed models; strategic fine-tuning shows that incentives can also shape the labels used before deployment. Work on learning from human feedback studies how human signals improve models; our framework studies the complementary case in which the feedback provider may prefer partial capability transfer. Data poisoning shows that adversarial data can harm learners; our credibility constraint studies a narrower, more economically motivated behavior in which reports remain close to the teacher's own beliefs.

We deliberately control the scope of the evidence. The synthetic environments make it possible to observe truth, beliefs, reports, and learner responses separately, and therefore to attribute differences in learner performance to strategic reporting rather than to variation in task difficulty. In contrast, real systems may involve complications like noisier incentives, multiple feedback providers, stronger auditing, and weaker strategic optimization. In particular, the human objective is a tractable proxy for preserving an advantage rather than a complete economic model of wages, substitution, or human values. The strategic teacher is also optimized directly; real humans may use simpler heuristics. Future work should connect the mechanism to richer economic and learning environments, including heterogeneous annotators and incentives created by repeated interaction.

Strategic fine-tuning, therefore, suggests an alignment-oriented auditing target: not just whether human reports are accurate or plausible, but whether the learner actually internalizes the information they contain. Potential mitigations that follow from the model include additional auditing near low-margin examples, training with multiple learner architectures to expose internalization gaps, and rewarding downstream generalization rather than immediate report plausibility alone. While these directions are not evaluated here, they follow from the same distinction between label quality, reporting incentives, and learner internalization. The core conclusion is that human-supervised pipelines depend on both the information in the labels and the incentives that generate them. When supervision transfers capability from agents who may face some form of substitution, even reports that remain close to belief can alter what is internalized by ordinary fine-tuning. Robust fine-tuning with human-generated data should therefore consider label quality, reporting incentives, and learner internalization together.

References

Battista Biggio, Blaine Nelson, and Pavel Laskov. Poisoning attacks against support vector machines. In *Proceedings of the 29th International Conference on Machine Learning*, 2012.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, 2017.

Vincent P. Crawford and Joel Sobel. Strategic information transmission. *Econometrica*, 50(6):1431–1451, 1982.

Zoë B. Cullen, Danielle Li, and Shengwu Li. Labor as Capital: AI and the Ownership of Expertise. Harvard Business School Working Paper No. 26-063, 2026.

Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. Strategic classification. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*, 2016.

Keegan Harris, Hoda Heidari, and Zhiwei Steven Wu. Stateful strategic regression. In *Advances in Neural Information Processing Systems*, 2021.

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. In *International Conference on Learning Representations*, 2022.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Safwan Hossain and Nisarg Shah. The effect of strategic noise in linear regression. *Autonomous Agents and Multi-Agent Systems*, 35(2):21, 2021.

Emir Kamenica and Matthew Gentzkow. Bayesian persuasion. *American Economic Review*, 101(6):2590–2615, 2011.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll 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. In *Advances in Neural Information Processing Systems*, 2022.

Juan C. Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

A Proofs

We collect the proofs for the formal statements in the main text. All expansions are local around an interior belief distribution, so the simplex constraint is handled by restricting perturbations δ to the tangent subspace satisfying $\sum_c \delta_c(x) = 0$ for each x.

A KL expansion. The following second-order approximation is used repeatedly. For two nearby interior distributions p and $p + u$ on the same simplex,

$$D_{\mathrm{KL}}(p \,\|\, p + u) = \frac{1}{2} u^\top \operatorname{diag}(p)^{-1} u + o(\|u\|^2),$$

where the first-order term vanishes because $\sum_c u_c = 0$. Averaging over x gives

$$\mathbb{E}_x D_{\mathrm{KL}}\big(b(\cdot \mid x) \,\|\, b(\cdot \mid x) + \delta(\cdot \mid x)\big) = \frac{1}{2}\|\delta\|_b^2 + o(\|\delta\|_b^2).$$

The same calculation also gives, for nearby perturbations δ and ζ, the expansion $D_{\mathrm{KL}}(b + \delta \,\|\, b + \zeta) = \frac{1}{2}\|\delta - \zeta\|_b^2 + o(\|\delta\|_b^2 + \|\zeta\|_b^2)$ after averaging over x.

Proof of Proposition 1. Fix the sequence of examples and reports observed by the learner. Once this sequence is fixed, the learner's objective at each update is exactly the supervised cross-entropy loss in Equation 1. The gradients, minibatches, optimizer states, and parameter updates are therefore identical to those produced by ordinary supervised fine-tuning on the same examples with the same soft labels. The strategic optimization that produced the reports no longer enters the learner's update.

If optimization is run to convergence over a model family $\mathcal{M}$, the learner minimizes

$$\mathbb{E}_x \mathrm{CE}\big(r(\cdot \mid x), p(\cdot \mid x)\big) = \mathbb{E}_x \mathrm{CE}\big(r(\cdot \mid x), r(\cdot \mid x)\big) + \mathbb{E}_x D_{\mathrm{KL}}\big(r(\cdot \mid x) \,\|\, p(\cdot \mid x)\big).$$

The first term is the entropy of the reports and does not depend on p. Thus every cross-entropy minimizer is also a minimizer of the expected KL divergence from the reports to the model distribution, giving the displayed characterization of $T(r)$. □

Proof of Theorem 1. Let $r = b + \delta$ with $\delta \in \mathcal{V}$. By the KL expansion above,

$$C(b + \delta; b) = \frac{1}{2}\|\delta\|_b^2 + o(\|\delta\|_b^2). \tag{5}$$

Differentiability of H along $\mathcal{V}$ gives the first-order expansion

$$H(b + \delta) - H(b) = \langle g_b, \delta \rangle_b + o(\|\delta\|_b). \tag{6}$$

The feasible set $C(b + \delta; b) \le \epsilon$ is therefore a Fisher ball of radius $\sqrt{2\epsilon}$ to first order. More explicitly, for any fixed $\eta > 0$ and sufficiently small ϵ, feasibility implies

$$\|\delta\|_b \le \sqrt{2\epsilon}\,(1 + \eta).$$

Using Cauchy-Schwarz in (6), every feasible perturbation satisfies

$$H(b + \delta) - H(b) \le \sqrt{2\epsilon}\,(1 + \eta)\|g_b\|_b + o(\sqrt{\epsilon}).$$

This gives the upper bound after sending $\eta \downarrow 0$.

For the matching lower bound, take

$$\delta_\epsilon = \sqrt{2\epsilon}\,\frac{g_b}{\|g_b\|_b}$$

when $g_b \neq 0$, with an arbitrarily small rescaling if needed to satisfy the constraint exactly. Substituting this direction into (5) and (6) yields

$$H(b + \delta_\epsilon) - H(b) = \sqrt{2\epsilon}\,\|g_b\|_b + o(\sqrt{\epsilon}).$$

If $g_b = 0$, the same display gives value $o(\sqrt{\epsilon})$. Combining the upper and lower bounds proves the theorem. □

Proof of Corollary 1. The calibrated assumptions give $T(b) = b$ and

$$T(b+\delta) = b + J_b[\delta] + o(\|\delta\|_b).$$

Thus the difference between the report and the post-training learner is

$$(b+\delta) - T(b+\delta) = (I - J_b)\delta + o(\|\delta\|_b) = M_b\delta + o(\|\delta\|_b).$$

Applying the local KL expansion to H gives

$$H(b+\delta) = \frac{1}{2}\|M_b\delta\|_b^2 + o(\|\delta\|_b^2). \tag{7}$$

The constraint expansion remains

$$C(b+\delta; b) = \frac{1}{2}\|\delta\|_b^2 + o(\|\delta\|_b^2).$$

The leading-order problem is therefore

$$\sup_{\delta\in\mathcal{V}} \frac{1}{2}\|M_b\delta\|_b^2 \quad \text{s.t.} \quad \frac{1}{2}\|\delta\|_b^2 \le \epsilon.$$

Writing $\delta = \sqrt{2\epsilon}\, u$ with $\|u\|_b = 1$, the leading value becomes

$$\epsilon \sup_{\substack{u\in\mathcal{V} \\ \|u\|_b=1}} \|M_b u\|_b^2 = \epsilon\rho_{\mathcal{V}}.$$

Because the adjoint is defined with respect to the Fisher inner product, this supremum is the top eigenvalue of $M_b^* M_b$ restricted to $\mathcal{V}$ when $\mathcal{V}$ is finite-dimensional. The value is positive exactly when there exists $u \in \mathcal{V}$ with $M_b u \neq 0$, which is precisely the failure of local internalization in an admissible direction. □

Proof of Proposition 2. The population response is the KL projection of a nearby report $b+\delta$ onto the smooth model family $\mathcal{M}$. Since $b \in \mathcal{M}$, any nearby model distribution can be written to first order as $b+\zeta$, where ζ lies in the tangent space $\mathcal{T}_b\mathcal{M}$. The local form of the projection objective is

$$\mathbb{E}_x D_{\mathrm{KL}}(b+\delta \,\|\, b+\zeta) = \frac{1}{2}\|\delta-\zeta\|_b^2 + o(\|\delta\|_b^2 + \|\zeta\|_b^2).$$

To first order, the projection therefore solves the least-squares problem

$$\min_{\zeta\in\mathcal{T}_b\mathcal{M}} \|\delta-\zeta\|_b^2.$$

The solution is the Fisher-orthogonal projection of δ onto $\mathcal{T}_b\mathcal{M}$. Hence the derivative of the response map is

$$J_b\delta = \mathrm{proj}_{\mathcal{T}_b\mathcal{M}}\,\delta,$$

and the mismatch operator is the normal residual

$$M_b\delta = \delta - J_b\delta = \mathrm{proj}_{(\mathcal{T}_b\mathcal{M})^\perp}\,\delta.$$

Substituting this expression into the definition of $\rho_{\mathcal{V}}$ gives the stated formula. □

Proof of Proposition 3. Let

$$a_b(x) = \arg\max_c b_c(x), \qquad a_r(x) = \arg\max_c r_c(x).$$

Suppose $m_b(x) > \gamma$ and $a_r(x) \neq a_b(x)$. Let $j = a_r(x)$. Since j is not the top class under b, we have

$$b_{a_b}(x) - b_j(x) > \gamma.$$

Since j is the top class under r, $r_j(x) \geq r_{a_b}(x)$. Therefore

$$\big(b_{a_b}(x) - r_{a_b}(x)\big) + \big(r_j(x) - b_j(x)\big) > \gamma.$$

The left-hand side is bounded above by

$$|b_{a_b}(x) - r_{a_b}(x)| + |r_j(x) - b_j(x)| \leq \|b(\cdot \mid x) - r(\cdot \mid x)\|_1.$$

Thus a change of the top class on a point with margin greater than γ implies

$$\mathrm{TV}\big(b(\cdot \mid x), r(\cdot \mid x)\big) \geq \gamma/2.$$

By Pinsker's inequality,

$$\mathrm{TV}\big(b(\cdot \mid x), r(\cdot \mid x)\big) \leq \sqrt{D_{\mathrm{KL}}(b(\cdot \mid x) \,\|\, r(\cdot \mid x))/2}.$$

Consequently, on points with margin greater than γ, a prediction change implies pointwise KL at least $\gamma^2/2$. Markov's inequality gives

$$\Pr_x[a_r(x) \neq a_b(x),\ m_b(x) > \gamma] \leq \frac{2}{\gamma^2}\mathbb{E}_x D_{\mathrm{KL}}\big(b(\cdot \mid x) \,\|\, r(\cdot \mid x)\big).$$

Adding the probability of the low-margin event $m_b(x) \leq \gamma$ proves the first claim.

For the accuracy bound, observe that whenever $a_r(x)$ is wrong and $a_b(x)$ is correct, the two predictions must differ. Hence

$$\Pr[a_r(x) \neq y^\star(x)] \leq \Pr[a_b(x) \neq y^\star(x)] + \Pr[a_r(x) \neq a_b(x)].$$

Rearranging and applying the first part yields the displayed lower bound on $\mathrm{Acc}(r)$. □