Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Screening Feedback for Language Models with Costly Verification

Zhonglin Liu

u3597461@connect.hku.hk

The University of Hong Kong

Jussi Keppo

National University of Singapore

Murari Mandal

Kalinga Institute of Industrial Technology

Hong Ming Tan

National University of Singapore

Article

Keywords: Costly verification, Incentives, User feedback, Quality control, Platform economics, Language models

Posted Date: May 6th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8771074/v2

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Screening Feedback for Language Models with Costly Verification

Zhonglin Liu[1*], Jussi Keppo[2], Murari Mandal[3], Hong Ming Tan[2]

[1*]Department of Mathematics, The University of Hong Kong, Pokfulam, Hong Kong, 000000, Hong Kong, China.
[2]Department of Analytics & Operations , National University of Singapore, 15 Kent Ridge Drive, Singapore, 119245, Singapore, Singapore.
[3]School of Computer Engineering, Kalinga Institute of Industrial Technology, Patia, Bhubaneswar, 751024, Odisha, India.

*Corresponding author(s). E-mail(s): u3597461@connect.hku.hk;
Contributing authors: keppo@nus.edu.sg; imurari.mandalfcs@kiit.ac.in;
thm@nus.edu.sg;

Abstract

Language model training and alignment rely on high-quality human feedback, yet platforms must incentivize valuable contributions while limiting harmful feedback from non-experts. We study a simple screening environment in which a platform commits to a uniform incentive policy $(\boldsymbol{\rho}, \boldsymbol{R}, \boldsymbol{P})$—a verification rate, a reward for submitting feedback, and a penalty imposed when verified feedback is harmful—and heterogeneous users decide whether to participate. High-type users are more likely to produce helpful feedback, while low-type users are more likely to generate harmful feedback.

We characterize the platform-optimal reward–penalty policy under costly verification in the robust pure-participation regimes. A key boundary condition, $\boldsymbol{\phi_H(1-\eta_H) = \phi_L(1-\eta_L)}$, separates parameter regions in which incentives implement normal separation (high types participate and low types abstain) from regions exhibiting reverse screening (low types participate while high types are deterred). Verification is the primary instrument: optimal policies feature full verification at moderate costs and selective verification as verification becomes more expensive. The optimal policy is typically penalty-constrained, with rewards pinned down by participation incentives and penalties limited by enforcement and reputational costs.

We further show that, under optimal verification, platform profit need not increase monotonically with population quality: profits can follow an inverted-U

pattern, peaking at intermediate shares of high-type users. The mechanism is that higher population quality induces the platform to reduce verification intensity, and the resulting cost savings may be insufficient to offset foregone verification benefits. Finally, we provide an illustrative simulation using a bigram language model as a transparent calibration exercise to generate plausible magnitudes for (η_H, η_L) and to visualize the model's comparative statics.

Keywords: Costly verification, Incentives, User feedback, Quality control, Platform economics, Language models

1 Introduction

Human feedback is a central input to modern language model training and alignment pipelines [1]. Yet platforms that solicit such feedback face a fundamental quality-control problem: how to attract users who are likely to provide helpful feedback while discouraging participation by users whose feedback is more likely to be harmful. This tension is particularly salient when user populations are heterogeneous, with a small set of experts coexisting with a larger mass of non-experts.

A distinctive feature of feedback for language models is that quality is often difficult to observe ex ante and costly to verify ex post. Unlike many crowdsourcing tasks with readily observable outcomes, assessing whether a feedback instance improves model behavior may require human review, offline evaluation, or additional experimentation. This creates an economic trade-off between stronger screening (more verification and harsher sanctions) and the operational, legal, and reputational costs of enforcing such policies.

We study this trade-off in a simple screening environment with costly verification. A platform commits to a uniform incentive policy (ρ, R, P) applied to all users: a verification rate ρ, a reward R for submitting feedback, and a penalty (sanction) P applied when verified feedback is harmful. Users differ in their likelihood of generating helpful feedback and in their effective sensitivity to penalties, and they decide whether to participate after observing the platform's policy.

Although we occasionally use the term "mechanism" in the broad platform-design sense, our model does *not* study a standard direct-revelation mechanism in which users report their types and the platform implements type-contingent allocations. Instead, we focus on the practically salient case in which expertise is difficult to certify and the platform's levers are verification and enforcement applied uniformly through product design and terms of service. In this sense, our setting is best viewed as a screening-and-enforcement problem with costly verification rather than a classical mechanism design problem.

Our analysis characterizes the platform-optimal reward–penalty policy under costly verification in the robust pure-participation regimes. A key boundary condition, $\phi_H(1-\eta_H) = \phi_L(1-\eta_L)$, partitions the parameter space into a region with normal separation (high types participate and low types abstain) and a region with reverse screening (low types participate while high types are deterred). Verification

is the primary instrument: optimal policies feature full verification at moderate costs and selective verification as verification becomes more expensive. The optimal policy is typically penalty-constrained: rewards are pinned down by participation incentives, while penalties are limited by enforcement and reputational costs, so deterrence and auditing jointly determine screening performance.

Because the platform commits to (ρ, R, P) before users choose whether to participate, non-degenerate mixed participation is a knife-edge phenomenon: it requires exact user indifference and is generically eliminated by arbitrarily small platform deviations that break indifference. Accordingly, we interpret $\phi_H(1-\eta_H) = \phi_L(1-\eta_L)$ as the boundary between the two robust pure-participation regimes (normal separation vs. reverse screening), rather than as a region with stable interior mixing.

We also provide comparative statics that highlight a potential non-monotonicity in platform profitability. Under optimal verification, profit need not increase monotonically with population quality: profits can follow an inverted-U pattern, peaking at intermediate shares of high-type users. The mechanism is that as the population becomes higher-quality, the platform optimally reduces verification intensity; the associated cost savings may be insufficient to offset foregone verification benefits.

To connect the theory to a transparent numerical example, we include an illustrative simulation based on a bigram language model trained on the Brown Corpus [2]. The purpose of this exercise is not to empirically validate the model, but to provide an interpretable calibration for the key heterogeneity parameters (η_H, η_L) and to visualize the paper's comparative statics in a controlled setting.

This work is most closely connected to research on screening and enforcement under costly verification, platform governance, and incentive provision for quality contributions [3–5]. We review and position these connections in Appendix A.

The remainder of the paper is organized as follows. Section 2 presents the model. Section 3 characterizes the optimal policy across regimes. Section 4 provides illustrative simulations. Section 5 concludes. Appendix A reviews related literature. Appendix B contains proofs.

2 Feedback Mechanism Design with Heterogeneous Users

We consider a game between a platform (a language-model provider) and a continuum of users indexed by $i \in [0, 1]$. Users can optionally provide feedback that may improve or harm the platform's model. The key friction is that assessing feedback quality is feasible but costly, so the platform chooses verification intensity and uses rewards and sanctions to screen participation.

The model is represented as a function $f : T^* \to \Delta(T)$ mapping an input token sequence to a distribution over the next token, following Duetting et al. [6]. Here T is the vocabulary, $\Delta(T)$ is the set of distributions over T, and $T^* = T \cup T^2 \cup \cdots \cup T^K$ is the set of token sequences up to context length K. Generation is autoregressive: starting with prompt $s_0 \in T^*$, the model samples $\tau_1 \sim f(s_0)$, forms $s_1 = s_0 \oplus \tau_1$, and iterates $\tau_k \sim f(s_{k-1})$, $s_k = s_{k-1} \oplus \tau_k$, until termination. The model f is parameterized

by architecture M and weights W, and platform quality depends on training and fine-tuning procedures that incorporate user feedback.

Each user i has a private type $t_i \in \{H, L\}$, where H denotes a high type (expert) and L denotes a low type (non-expert). Types are i.i.d. with $\mathbb{P}(t_i = H) = \lambda$ and $\mathbb{P}(t_i = L) = 1 - \lambda$.

If a user provides feedback, it can be helpful or harmful. We classify a feedback instance as helpful if it increases expected log-likelihood on a held-out validation distribution. Formally, let $r_i := \mathbf{1}\{\tilde{\sigma}(s_i, \tau_i) > 0\}$, where

$$\tilde{\sigma}(s_i, \tau_i) := \mathbb{E}_{(s', \tau') \sim \mathcal{V}} \left[\log f_{T+1}^{(i)}(\tau' \mid s') - \log f_T(\tau' \mid s') \right].$$

This likelihood-improvement criterion is motivated by recent likelihood-based views of alignment from preference feedback. In particular, Tang and Feng [7] propose Ranked Choice Preference Optimization (RCPO), which frames preference alignment as maximum likelihood estimation of (ranked) choice models. In RCPO, the training objective is a (ranked) choice log-likelihood $\sum_i \log g(\mu_i^k, S_i, \{r_{\pi_\theta}(x_i, y)\}_{y \in S_i})$, where the reward is derived from policy log-probability ratios in the DPO family, e.g.,

$$r_{\pi_\theta}(x, y) = \beta \log \frac{\pi_\theta(y \mid x)}{\pi_{\text{ref}}(y \mid x)} + \beta \log Z(x),$$

so that updates are naturally driven by log-probability (log-likelihood) differences. Under this perspective, using held-out expected log-likelihood improvement $\tilde{\sigma}(s_i, \tau_i)$ as a proxy for whether incorporating a feedback instance improves model fit provides a principled, likelihood-consistent notion of "helpfulness".

We define the type-dependent helpfulness probability

$$\eta_t := \mathbb{P}(r_i = 1 \mid t_i = t) = \mathbb{P}\big(\tilde{\sigma}(s_i, \tau_i) > 0 \mid t_i = t\big),$$

and assume $\eta_H > \eta_L$.

The interaction is a sequential (Stackelberg) game:

1. The platform commits to a uniform policy (ρ, R, P), where $\rho \in [0, 1]$ is the verification rate, $R \geq 0$ is the reward for submitting feedback, and $P \geq 0$ is a penalty (sanction) applied when verified feedback is harmful.
2. Users observe (ρ, R, P) and each user chooses an action $a_i \in \{F, N\}$, where F denotes providing feedback (incurring cost $c > 0$) and N denotes not providing feedback.
3. If $a_i = F$, the user submits a feedback string $\tau_i \in T^*$, yielding a feedback instance (s_i, τ_i). The platform verifies the instance with probability ρ: let $v_i \sim \text{Bernoulli}(\rho)$. If $v_i = 1$, the platform observes whether the feedback is harmful ($r_i = 0$) and applies the penalty accordingly.

Verification incurs a per-feedback cost $k > 0$. Transfers to user i are

$$m_i = \begin{cases} R & \text{if } a_i = F \text{ and } (v_i = 0 \text{ or } r_i = 1), \\ R - P & \text{if } a_i = F \text{ and } v_i = 1 \text{ and } r_i = 0, \\ 0 & \text{if } a_i = N. \end{cases}$$

Equivalently, $m_i = \mathbf{1}\{a_i = F\}\,[R - P \cdot \mathbf{1}\{v_i = 1, r_i = 0\}]$. In practice, P need not be a monetary fine: it can be implemented through escrow/deposits that are forfeited upon verified harmful feedback, reward clawbacks or withholding of future payouts, or account-level sanctions (e.g., suspension from the feedback program) with an equivalent monetized cost. Accordingly, we interpret P as a reduced-form measure of expected sanction severity conditional on verified harmful feedback, while allowing for enforcement and reputational constraints captured by the platform's penalty cost term below.

For unverified feedback ($v_i = 0$), the platform does not observe whether the instance is helpful or harmful. Let η_u^+ be the probability that an unverified feedback instance is helpful and η_u^- the probability it is harmful, with $\eta_u^+, \eta_u^- \in [0, 1]$ and $\eta_u^+ + \eta_u^- = 1$. The term $(\eta_u^+ - \eta_u^-)$ captures the net expected impact of unverified feedback.

User decision-making is modeled using a reduced-form utility derived from mean–variance preferences. The mean–variance formulation includes variance aversion with type-specific parameters $\psi_t \geq 0$, but the resulting quadratic terms complicate the equilibrium analysis. We therefore adopt a local approximation (Appendix B.1) that preserves the endogenous incentive term while absorbing higher-order risk adjustments into a constant type-specific coefficient. This approach parallels perturbation-based approximations used to recover tractable incentives under risk aversion (see, e.g., Fibich et al. [8]).

Let $\bar{U} \equiv U(s_{i,k})$ denote the deterministic base utility from receiving the model response (independent of type). Under the approximation, type-t utility is

$$u_t(a_i) = \begin{cases} \bar{U} + R - c - \phi_t P q_t & \text{if } a_i = F, \\ \bar{U} & \text{if } a_i = N, \end{cases}$$

where $q_t \equiv \rho(1 - \eta_t)$ is the probability a type-t user faces a penalty when providing feedback, and $\phi_t \equiv 1 + \psi_t \bar{P}(1 - \bar{q}_t)$ is a type-specific risk-adjusted penalty coefficient defined around a reference point $(\bar{P}, \bar{q}_t)$. Thus, a type-t user provides feedback if and only if

$$R - c \geq \phi_t P q_t = \phi_t P \rho(1 - \eta_t).$$

Let $\theta^{F,H}$ and $\theta^{F,L}$ denote the mass of feedback supplied by high and low types. Verified and unverified masses are

$$\theta^{v,H} = \rho\,\theta^{F,H}, \quad \theta^{v,L} = \rho\,\theta^{F,L}, \quad \theta^{u} = (1 - \rho)(\theta^{F,H} + \theta^{F,L}).$$

Model quality improvement is

$$\Delta Q = \gamma_v\left(\eta_H\theta^{v,H} + \eta_L\theta^{v,L}\right) + \gamma_u(\eta_u^+ - \eta_u^-)\theta^u,$$

where $\gamma_v > \gamma_u > 0$ weight verified and unverified feedback.

The platform chooses (ρ, R, P) to maximize profit, trading off quality improvements, verification costs, transfers, and enforcement/reputational costs. We assume a quadratic verification cost $\frac{k}{2}(\theta^{v,H} + \theta^{v,L})^2$ to capture diseconomies of scale in review, and a quadratic reputational/enforcement cost $\alpha P^2(\theta^{F,H} + \theta^{F,L})$ to reflect increasing marginal costs of imposing harsher sanctions. The platform's objective is

$$\begin{aligned}\Pi(\rho, R, P) = &\underbrace{\gamma_v[\eta_H\theta^{v,H} + \eta_L\theta^{v,L}] + \gamma_u(\eta_u^+ - \eta_u^-)\theta^u}_{\text{Quality improvement}} - \underbrace{\frac{k}{2}(\theta^{v,H} + \theta^{v,L})^2}_{\text{Verification costs}} \\ &- \underbrace{R(\theta^{F,H} + \theta^{F,L})}_{\text{Reward payments}} + \underbrace{P\rho\left[(1-\eta_H)\theta^{F,H} + (1-\eta_L)\theta^{F,L}\right]}_{\text{Penalty collections}} - \underbrace{\alpha P^2(\theta^{F,H} + \theta^{F,L})}_{\text{Reputational/enforcement costs}}.\end{aligned} \tag{1}$$

We focus on subgame-perfect equilibrium outcomes of this sequential game given the platform's commitment to (ρ, R, P).

3 Results

3.1 Perfect Separation

Theorem 1 (Perfect Separation Condition) *Perfect separation equilibrium, where high-types provide feedback ($\theta^{F,H} = \lambda$) and low-types abstain ($\theta^{F,L} = 0$), exists if and only if the following condition holds:*

$$\phi_H(1-\eta_H) < \phi_L(1-\eta_L) \tag{2}$$

Equivalently, this can be expressed as:

$$\frac{\phi_H}{\phi_L} < \frac{1-\eta_L}{1-\eta_H} \tag{3}$$

Moreover, in any perfect separation equilibrium, the reward level must satisfy:

$$R = c + \phi_H P\rho(1-\eta_H) \tag{4}$$

Proof See Appendix B.2. □

Economic Implications: This theorem establishes the basic screening feasibility condition for AI training platforms, revealing that successful separation requires behavioral heterogeneity to exceed quality heterogeneity. The inequality $\frac{\phi_H}{\phi_L} < \frac{1-\eta_L}{1-\eta_H}$ implies that platforms achieve first-best outcomes when low-quality users exhibit sufficiently higher penalty sensitivity (ϕ_L) to offset their higher error rates (since $1-\eta_L > 1-\eta_H$, low-types have higher probability of harmful feedback). The reward equation $R = c + \phi_H P\rho(1-\eta_H)$ demonstrates that penalties are the fundamental

screening instrument: when $P = 0$, rewards reduce to the participation cost $R = c$, making both types indifferent and preventing separation, whereas positive penalties create type-dependent expected costs $\phi_t P\rho(1-\eta_t)$ that enable differentiation through heterogeneous risk perception $\phi_H \neq \phi_L$. The mechanism exhibits incentive compatibility where high-types are indifferent between participating and abstaining while low-types strictly prefer abstention.

From a welfare perspective, perfect separation maximizes platform value by eliminating harmful feedback while maintaining expert participation, though it may reduce total user welfare by excluding low-types. For platform design, understanding penalty sensitivity differences is as important as measuring feedback quality differences for effective screening mechanisms.

Theorem 2 (Optimal Mechanism under Perfect Separation) *Under perfect separation conditions, the platform's optimal mechanism is characterized by the following two cases:*

Case 1: $\phi_H < 1$***,*** *when the condition*

$$k\lambda > \frac{(1-\eta_H)^2(1-\phi_H)^2}{2\alpha} \tag{5}$$

is satisfied, the optimal mechanism is:

$$\rho^* = \min\left\{1, \max\left\{0, \frac{\gamma_v\eta_H - \gamma_u(\eta_u^+ - \eta_u^-)}{k\lambda - \frac{(1-\eta_H)^2(1-\phi_H)^2}{2\alpha}}\right\}\right\} \tag{6}$$

$$P^* = \frac{\rho^*(1-\eta_H)(1-\phi_H)}{2\alpha} \tag{7}$$

$$R^* = c + \phi_H P^*\rho^*(1-\eta_H) \tag{8}$$

Case 2: *When $\phi_H \geq 1$, or when $\phi_H < 1$ but condition (5) is violated, the optimal mechanism is characterized as follows:*

(i) If $\phi_H \geq 1$***:***

$$\rho^* = \min\left\{1, \max\left\{0, \frac{\gamma_v\eta_H - \gamma_u(\eta_u^+ - \eta_u^-)}{k\lambda}\right\}\right\} \tag{9}$$

$$P^* = 0 \tag{10}$$

$$R^* = c \tag{11}$$

(ii) If $\phi_H < 1$ ***and*** (5) ***is violated:***

$$\rho^* \in \{0, 1\}, \tag{12}$$

with the boundary choice given by

$$\rho^* = \begin{cases} 1, & \text{if } \gamma_v\eta_H - \gamma_u(\eta_u^+ - \eta_u^-) - \dfrac{k\lambda}{2} + \dfrac{(1-\eta_H)^2(1-\phi_H)^2}{4\alpha} \geq 0, \\ 0, & \text{otherwise,} \end{cases} \tag{13}$$

and

$$P^* = \frac{\rho^*(1-\eta_H)(1-\phi_H)}{2\alpha}, \tag{14}$$

$$R^* = c + \phi_H P^*\rho^*(1-\eta_H). \tag{15}$$

3.2 Mixed Strategies (Non-Robust under Sequential Commitment)

Our model is a sequential (Stackelberg) game: the platform commits to a policy (ρ, R, P) and users then decide whether to provide feedback. In such commitment environments, non-degenerate mixed participation (i.e., $p_t \in (0,1)$ for some type $t \in \{H, L\}$) is a knife-edge phenomenon. It relies on exact user indifference and can be eliminated by an arbitrarily small platform deviation. Consequently, non-degenerate mixing does not arise in *generic platform-optimal* subgame-perfect equilibrium outcomes under commitment, and we focus on the two robust regimes: perfect separation and reverse screening.

Let a type-t user's expected payoff gain from providing feedback (relative to not providing feedback) be

$$\Delta_t(\rho, R, P) \;\equiv\; (R - c) \;-\; \phi_t \, P \, \rho \, (1 - \eta_t), \qquad t \in \{H, L\}. \tag{16}$$

Then type t chooses F if $\Delta_t > 0$, chooses N if $\Delta_t < 0$, and is indifferent if $\Delta_t = 0$.

Proposition 3 (Non-degenerate mixing is not robust in generic SPNE) *Fix any (ρ, R, P). If $\Delta_t(\rho, R, P) \neq 0$, then type t has a unique (pure) best response, hence $p_t \in \{0, 1\}$. If $\Delta_t(\rho, R, P) = 0$, then type t is indifferent and any $p_t \in [0, 1]$ is possible.*

Moreover, if $\Delta_t(\rho, R, P) = 0$ for some type t, then for any $\varepsilon > 0$ the platform can perturb the reward by an arbitrarily small amount, setting $R' = R + \varepsilon$ or $R' = R - \varepsilon$, to obtain $\Delta_t(\rho, R', P) \neq 0$ and thereby induce a pure *best response ($p'_t = 1$ or $p'_t = 0$). Hence any candidate equilibrium outcome with $p_t \in (0, 1)$ is non-robust: unless the platform is exactly indifferent between the two induced pure outcomes, an arbitrarily small deviation breaks the user's indifference and eliminates mixing. Equivalently, any interior-mixing outcome can persist only on a knife-edge where the platform is exactly indifferent between the neighboring pure-participation outcomes induced by such perturbations.*

Proof The user's Stage-2 objective is linear in the action (or mixing probability), so a strict sign of Δ_t implies a unique pure best response. If $\Delta_t = 0$, the user is indifferent.

If $\Delta_t(\rho, R, P) = 0$, then $\Delta_t(\rho, R + \varepsilon, P) = \varepsilon > 0$ and $\Delta_t(\rho, R - \varepsilon, P) = -\varepsilon < 0$. Thus an arbitrarily small perturbation of R breaks indifference and yields a pure best response. (An equivalent argument applies by perturbing ρ or P by an arbitrarily small amount whenever these instruments are continuous choices.) In a Stackelberg game, if the platform strictly prefers one of these induced pure outcomes, it can choose the corresponding sign of ε to improve profit, contradicting optimality of any policy that relies on interior mixing. □

Thus the "mixed" condition is a boundary, not a robust regime. The equality

$$\phi_H(1 - \eta_H) = \phi_L(1 - \eta_L) \tag{17}$$

implies that the two types have the same expected (risk-adjusted) penalty exposure per unit $P\rho$. In particular, any policy satisfying the high-type participation indifference $R = c + \phi_H P \rho (1 - \eta_H)$ also makes the low type indifferent (and vice versa), so

strict type sorting via (R, P, ρ) is impossible on (17). However, by Lemma 3, interior mixing at such indifference points is not robust under commitment: arbitrarily small platform deviations generically select pure participation.

Accordingly, we do not treat mixed or semi-separating (one-type-mixing) outcomes as separate equilibrium regimes in the sequential model. Instead, (17) is best interpreted as the knife-edge boundary separating the two robust screening regimes:

- **Perfect separation** when $\phi_H(1-\eta_H) < \phi_L(1-\eta_L)$ (Theorem 1);
- **Reverse screening** when $\phi_H(1-\eta_H) > \phi_L(1-\eta_L)$ (Theorem 4).

3.3 Reverse Screening

A particularly important result emerges when penalty-based mechanisms produce sorting effects opposite to their intended design. We formalize this phenomenon through the following theorems:

Theorem 4 (Reverse Screening Conditions (Strict)) *Reverse screening equilibria with $(p_H, p_L) = (0, 1)$ i.e., low-quality users participate and high quality users are strictly deterred exist if and only if:*

$$\phi_L(1-\eta_L) < \phi_H(1-\eta_H). \tag{18}$$

Equivalently,

$$\frac{\phi_H}{\phi_L} > \frac{1-\eta_L}{1-\eta_H}. \tag{19}$$

Moreover, in any optimal reverse-screening mechanism the low-type participation constraint binds, so the reward satisfies

$$R \;=\; c + \phi_L\, P\, \rho\, (1-\eta_L). \tag{20}$$

Proof The proof can be found in Appendix B.4. □

Theorem 5 (Optimal Mechanism under Reverse Screening) *Suppose the reverse-screening condition in Theorem 4 holds:*

$$\phi_L(1-\eta_L) \le \phi_H(1-\eta_H). \tag{21}$$

Consider reverse screening, i.e., low-quality users participate and high-quality users are deterred ($p_L = 1$, $p_H = 0$). Then the platform's optimal mechanism is characterized by the following two cases.

Case 1: $\phi_L < 1$, *when the condition*

$$k(1-\lambda) > \frac{(1-\eta_L)^2(1-\phi_L)^2}{2\alpha} \tag{22}$$

is satisfied, the optimal mechanism is:

$$\rho^* = \min\left\{1, \max\left\{0, \frac{\gamma_v\eta_L - \gamma_u(\eta_u^+ - \eta_u^-)}{k(1-\lambda) - \frac{(1-\eta_L)^2(1-\phi_L)^2}{2\alpha}}\right\}\right\}, \tag{23}$$

$$P^* = \frac{\rho^*(1-\eta_L)(1-\phi_L)}{2\alpha}, \tag{24}$$

$$R^* = c + \phi_L P^* \rho^* (1-\eta_L). \tag{25}$$

Case 2: *When $\phi_L \geq 1$, or when $\phi_L < 1$ but condition (22) is violated, the optimal mechanism is characterized as follows:*

(i) If $\phi_L \geq 1$***:***

$$\rho^* = \min\left\{1, \max\left\{0, \frac{\gamma_v \eta_L - \gamma_u(\eta_u^+ - \eta_u^-)}{k(1-\lambda)}\right\}\right\}, \tag{26}$$

$$P^* = 0, \tag{27}$$

$$R^* = c; \tag{28}$$

(ii) If $\phi_L < 1$ ***and*** (22) ***is violated:***

$$\rho^* \in \{0, 1\}, \tag{29}$$

with the boundary choice given by

$$\rho^* = \begin{cases} 1, & \text{if } \gamma_v \eta_L - \gamma_u(\eta_u^+ - \eta_u^-) - \dfrac{k(1-\lambda)}{2} + \dfrac{(1-\eta_L)^2(1-\phi_L)^2}{4\alpha} \geq 0, \\ 0, & \text{otherwise}, \end{cases} \tag{30}$$

and

$$P^* = \frac{\rho^*(1-\eta_L)(1-\phi_L)}{2\alpha}, \tag{31}$$

$$R^* = c + \phi_L P^* \rho^*(1-\eta_L). \tag{32}$$

Implementation of reverse screening. *If $\phi_L(1-\eta_L) < \phi_H(1-\eta_H)$, then the mechanism above strictly deters high types whenever the low-type participation constraint binds. If instead $\phi_L(1-\eta_L) = \phi_H(1-\eta_H)$, then high types are (at most) indifferent at the same cutoff; in that knife-edge case, reverse screening requires an equilibrium selection (e.g., tie-breaking against participation for the high type), consistent with the non-robustness of mixing under commitment.*

Proof The proof can be found in Appendix B.5. □

4 Illustrative Simulations and Stylized Calibration

Our theoretical results characterize the platform-optimal verification–reward–penalty policy across screening regimes. This section provides a transparent numerical illustration of the model's primitives and comparative statics. Because the simulation environment is intentionally simple, it is not intended as an empirical validation of the model or as a quantitative claim about large language model training. Rather, it serves two purposes: (i) to produce plausible magnitudes for the key heterogeneity parameters (η_H, η_L) in a controlled setting, and (ii) to visualize how optimal verification and profit respond to changes in costs and population composition.

We conduct two complementary exercises. A micro-simulation uses a bigram language model trained on the Brown Corpus [2] to obtain a stylized calibration of the helpfulness probabilities η_H and η_L. A macro-simulation then uses these calibrated magnitudes to illustrate the platform's optimal policy and the implied profit comparative statics across population shares and cost parameters.

4.1 Micro-Simulation for Stylized Calibration

To obtain transparent magnitudes for (η_H, η_L), we implement a micro-simulation using a bigram language model trained on the Brown Corpus. The bigram model is a minimal and interpretable testbed in which we can explicitly control the source of model errors and the quality of user feedback. The goal is not to represent the full complexity of modern language models, but to generate a simple environment in which high-type feedback is more likely to be helpful than low-type feedback, consistent with the model assumption $\eta_H > \eta_L$.

We estimate η_H and η_L using 10,000 Monte Carlo iterations.

4.1.1 Simulation Setup

We train a bigram language model f_{truth} on 2,000 sentences from the Brown Corpus with a vocabulary of 5,000 tokens. The model represents transition probabilities $P(w_{t+1} \mid w_t)$ estimated via maximum likelihood with Laplace smoothing ($\alpha = 0.01$). To create a model that can benefit from feedback, we generate a flawed model f_T by intentionally degrading 100 randomly selected prompts. For each flawed prompt w_t, we identify the correct continuation w^* (highest probability under f_{truth}) and an alternative w^- (randomly selected from the top five alternatives), then swap and exaggerate these probabilities:

$$f_T(w^* \mid w_t) = 0.5 \cdot f_{\text{truth}}(w^- \mid w_t), \qquad f_T(w^- \mid w_t) = 1.5 \cdot f_{\text{truth}}(w^* \mid w_t), \tag{33}$$

followed by renormalization.

We model two user types with different feedback accuracies: High-Type users provide correct feedback with probability 88%, while Low-Type users provide correct feedback with probability 38%. Upon receiving a feedback token $\hat{w}$ for prompt w_t, we update the model using an additive learning rule with rate $\delta = 0.2$, and we evaluate feedback helpfulness using expected log-likelihood improvement on a validation corpus:

$$\tilde{\sigma} = \mathbb{E}_{(w_t, w_{t+1}) \sim \text{Validation}} \left[\log f_{T+1}(w_{t+1} \mid w_t) - \log f_T(w_{t+1} \mid w_t) \right]. \tag{34}$$

Feedback is classified as helpful if $\tilde{\sigma} > 0$.

4.1.2 Results

Table 1 presents example interactions. The table includes prompts with [UNK] (unknown) tokens, representing words outside the training vocabulary or tokens mapped to the unknown category during preprocessing. High-Type users provide correct feedback more frequently and more often generate positive $\tilde{\sigma}$, while Low-Type users more often generate negative $\tilde{\sigma}$.

Table 2 summarizes results from 10,000 Monte Carlo iterations. The estimated helpfulness probability for High-Type users is $\eta_H = 0.892$. For Low-Type users, we obtain $\eta_L = 0.427$, which is higher than the 38% feedback accuracy because incorrect feedback can occasionally improve validation performance by chance. The Low-Type

Table 1 Example user interactions showing differential feedback quality between High-Type (experts) and Low-Type (non-experts) users. [UNK] denotes unknown tokens not in the model's vocabulary.

Prompt	Truth	Model	H-Feed	H-Correct	H-$\tilde{\sigma}$	L-Feed	L-Correct	L-$\tilde{\sigma}$
out	of	in	of	Yes	+8.4e-4	of	Yes	+8.4e-4
motel	in	issues	in	Yes	+5.0e-4	issues	No	-2.0e-5
enough	to	.	to	Yes	+1.8e-4	.	No	-7.7e-6
who	[UNK]	are	[UNK]	Yes	+3.3e-4	[UNK]	Yes	+3.3e-4
–	the	and	the	Yes	+4.3e-4	and	No	-3.7e-4
pink	[UNK]	awake	[UNK]	Yes	+5.0e-4	awake	No	-1.6e-5
care	plan	.	.	No	-1.9e-5	.	No	-1.9e-5
f.	[UNK]	lee	[UNK]	Yes	+4.6e-4	lee	No	-3.1e-5

Table 2 Summary statistics from 10,000 Monte Carlo simulations.

Metric	High-Type	Low-Type
Helpfulness probability (η)	0.892	0.427
Mean $\tilde{\sigma}$	7.70×10^{-4}	2.05×10^{-4}
Median $\tilde{\sigma}$	2.58×10^{-4}	-6.86×10^{-6}
Std. dev. $\tilde{\sigma}$	4.24×10^{-3}	2.85×10^{-3}
Ratio η_H/η_L	2.09	
Difference $\eta_H - \eta_L$	0.465	

median $\tilde{\sigma}$ is negative, indicating that more than half of non-expert feedback is harmful in this stylized environment.

4.2 Macro-Simulation: Optimal Policy and Comparative Statics

We next illustrate the platform's optimal policy and resulting profits as we vary key economic parameters and the population share of high types. The macro-simulation uses parameter magnitudes consistent with Section 4.1 and is intended to visualize regime boundaries and comparative statics rather than to provide quantitative predictions.

4.2.1 Parameter Configuration

We set the helpfulness probabilities to $\eta_H = 0.88$ and $\eta_L = 0.38$, preserving the large quality gap observed in the micro-simulation while using a slightly more conservative η_L to reflect additional noise and heterogeneity not captured by the bigram testbed. Platform benefits are $\gamma_v = 2.2$ for verified feedback. We set the net quality impact from unverified feedback, $\gamma_u(\eta_u^+ - \eta_u^-)$, to 0.4, reflecting that unverified feedback is on average helpful but less valuable than verified feedback.

The verification cost parameter is $k = 3.0$, and reputation/enforcement sensitivity is $\alpha = 0.8$. Penalty sensitivities are $\phi_H = 0.5$ and $\phi_L = 1.2$, capturing the possibility that lower-quality users are effectively more penalty-sensitive. The baseline population share is $\lambda = 0.25$, and the participation cost is $c = 0.5$.

4.2.2 Equilibrium Regimes and Verification Strategy

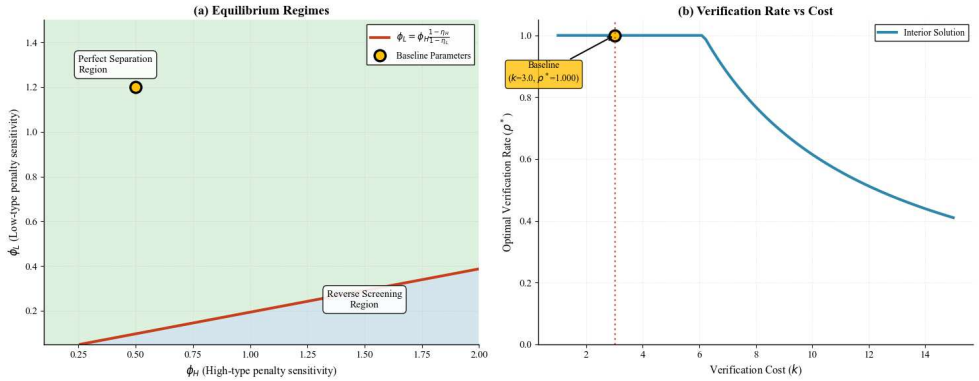


Fig. 1 Equilibrium regimes and verification strategy. Panel (a) partitions the parameter space by the boundary condition $\phi_H(1-\eta_H) = \phi_L(1-\eta_L)$ into normal separation and reverse screening. Panel (b) illustrates the optimal verification rate ρ^* as verification cost k varies.

Figure 1 illustrates two regime regions separated by $\phi_H(1-\eta_H) = \phi_L(1-\eta_L)$. In the normal-separation region, $\phi_H(1-\eta_H) < \phi_L(1-\eta_L)$, the platform can deter low types while maintaining high-type participation. In the reverse-screening region, $\phi_H(1-\eta_H) > \phi_L(1-\eta_L)$, policies that make low types willing to participate can deter high types.

Panel (b) illustrates how optimal verification responds to verification costs. At low costs, the platform chooses high verification intensity; as k increases, optimal verification becomes selective. The figure visualizes the threshold behavior implied by the theoretical characterizations.

4.2.3 Population Quality and the Inverted-U Profit Relationship

Figure 2 illustrates an inverted-U relationship between population quality and platform profit: profit initially rises with λ, peaks around $\lambda \approx 0.45$ under this calibration, and then declines for very high-quality populations. This pattern is consistent with the non-monotonicity result formalized below.

Proposition 6 (Non-Monotonicity of Profit in Population Quality) *Maintain the perfect-separation equilibrium of Theorem 2. Suppose $\phi_H < 1$ and the interior condition holds:*

$$k\lambda > \frac{(1-\eta_H)^2(1-\phi_H)^2}{2\alpha}. \tag{35}$$

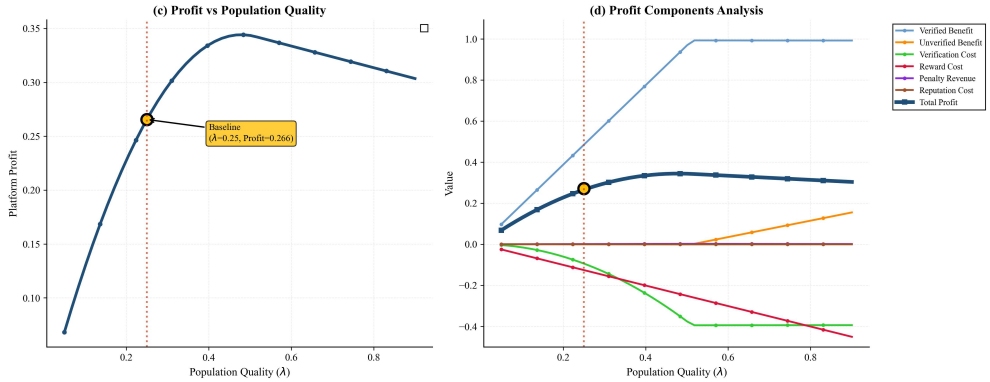


Fig. 2 Population Quality and Profit Components Analysis. Panel (c) shows the relationship between population quality λ and platform profit. Panel (d) decomposes this relationship into constituent components. Note: reputation cost and penalty collections appear negligible due to the small optimal penalty levels under our parameter values, though they are non-zero.

Define

$$N \equiv \gamma_v \eta_H - \gamma_u(\eta_u^+ - \eta_u^-), \qquad D \equiv \frac{(1-\eta_H)^2(1-\phi_H)^2}{2\alpha}. \tag{36}$$

If the participation cost satisfies

$$c > \gamma_u(\eta_u^+ - \eta_u^-) - \frac{N^2 D}{2(k-D)^2}, \tag{37}$$

then the platform's optimal profit $\Pi^*(\lambda)$ *is not monotonically increasing in* $\lambda \in (0,1)$*; in particular, there exists* $\bar{\lambda} \in (0,1)$ *such that*

$$\left.\frac{d\Pi^*}{d\lambda}\right|_{\lambda=\bar{\lambda}} < 0. \tag{38}$$

Proof See Appendix B.6. □

Panel (d) decomposes profit into its constituent components. Verified benefits initially increase with λ but flatten as the platform optimally adjusts verification. Verification costs and reward costs decline with λ because higher population quality makes harmful feedback less likely and therefore reduces the marginal value of verification. Under the binding participation condition in the separating region, the reward is tied to the penalty and verification choice, so changes in (P^*, ρ^*) mechanically translate into changes in reward expenditures.

The decline in profits at high λ stems from an optimal but countervailing verification response. As λ increases, the platform reduces ρ^*; while this saves verification and reward costs, it can also reduce the realized benefits from verified feedback. Under the conditions in Proposition 6, the latter effect dominates over some range, yielding non-monotonic profit in population quality.

5 Conclusion

This paper studies incentives for user feedback under costly verification in language model training environments. We analyze a platform that commits to a uniform policy (ρ, R, P)—a verification rate, a reward for submitting feedback, and a penalty (sanction) imposed when verified feedback is harmful—and a heterogeneous user population that differs both in feedback quality and in effective penalty sensitivity.

Our main results characterize when screening is feasible and how optimal verification and sanctions interact. Successful screening depends on behavioral heterogeneity exceeding quality differences: the platform can implement normal separation (high types participate and low types abstain) when

$$\frac{\phi_H}{\phi_L} < \frac{1-\eta_L}{1-\eta_H},$$

and it enters a reverse-screening region when this condition fails, in which policies that induce low-type participation can deter high types. This reverse-screening regime is a failure mode for penalty-based governance: stronger sanctions need not improve outcomes if they disproportionately discourage high-quality contributors.

Across regimes, verification is the primary strategic instrument. When verification costs are moderate, the platform optimally chooses high verification intensity; as verification becomes more expensive, optimal policies shift toward selective verification. The optimal policy is typically penalty-constrained: rewards are pinned down by participation incentives, while penalties are limited by enforcement and reputational considerations, so screening performance is determined jointly by auditing and deterrence.

We also highlight a non-monotonic profit comparative static. Under optimal verification, platform profit need not increase monotonically with population quality; it can exhibit an inverted-U pattern, peaking at intermediate shares of high-type users. The mechanism is that as the population becomes higher-quality, the platform optimally reduces verification intensity, and the resulting cost savings may be insufficient to offset foregone verification benefits.

Our simulations are intended as a transparent illustration rather than an empirical validation. Using a bigram language model testbed, we provide a stylized calibration that generates plausible magnitudes for (η_H, η_L) and visualizes the model's comparative statics.

Several limitations suggest directions for future work. We assume static types and a simple verification technology, abstract from learning about user quality over time and from platform competition, and represent enforcement constraints in reduced form. Extending the model to dynamic settings with reputation, Bayesian learning, multi-platform interaction, and endogenous investments in verification would strengthen its applicability. Empirical work using operational data from feedback programs could further discipline parameter choices and quantify the impact of alternative verification and enforcement designs.

6 Declarations

Competing Interests

The authors declare that they have no competing interests.

Funding

No funding was received for this work.

Author Contributions

L.Z. conceived and designed the experiments; performed the experiments; analyzed the data; contributed materials/analysis tools; and wrote the paper. H.M.T. conceived and designed the experiments; contributed materials/analysis tools; and wrote the paper. J.K. and M.M. contributed materials/analysis tools. All authors reviewed and approved the final manuscript.

Appendix A Related Work

Our paper is related to work on platform governance, incentives for quality provision, and screening under costly verification. We study a platform that commits to a uniform policy (ρ, R, P) with endogenous verification and enforceable sanctions; we do not analyze direct-revelation mechanisms with type reports and menu-based allocations. This positioning determines which literatures are most directly comparable.

Costly verification, auditing, and enforcement. A central feature of our model is that feedback quality can be assessed through verification, but only at a cost, and the platform strategically chooses verification intensity. This connects our setting to classic contract-theoretic and principal–agent models with costly state verification and auditing [3–5]. Our contribution is to characterize how endogenous verification interacts with reward–penalty incentives to generate distinct screening regimes (normal separation versus reverse screening) and to deliver comparative statics for optimal verification and sanctions in a feedback-quality setting.

Platforms and governance instruments. Our analysis also relates to platform economics, which emphasizes how platforms shape participation and value creation through prices and rules [9, 10]. Dynamic feedback and learning considerations in platform environments are studied by Cabral and Hortacsu [11]. Relative to this literature, we focus on a specific governance toolkit—probabilistic verification and sanctions—and study how these instruments screen heterogeneous contributors when feedback quality is imperfectly observed.

Crowdsourcing, quality elicitation, and peer prediction. A large literature studies how to incentivize informative reports and high-quality contributions when direct verification is difficult. Peer prediction and related methods [12, 13] use statistical relationships across reports to elicit information without ground truth, and related work examines endogenous proficiency and information aggregation in crowdsourcing [14, 15]. These approaches are complementary to ours: whereas peer prediction substitutes for verification, our model treats verification as feasible but costly and asks

how an optimizing platform should combine probabilistic verification with rewards and penalties to screen participation by heterogeneous users.

Data quality in ML pipelines and AI-related incentive design. In machine learning systems, data quality is often addressed with technical tools such as data cleaning and uncertainty-based selection [16, 17]. Economic mechanisms for procuring data appear in adjacent settings such as mobile crowdsensing [18], while reputation and repeated-interaction systems provide alternative ways to identify high-quality contributors [19, 20]. Recent work also explores incentive design in AI training contexts [6, 21, 22]. Our focus differs by centering costly verification and enforceable sanctions as the key governance instruments when expertise is difficult to certify and historical performance signals are limited.

Appendix B Proofs of Main Results

B.1 Constant Risk Adjustment Approximation

Proposition 7 (Constant risk adjustment approximation) *Let $q_t \equiv \rho(1-\eta_t) \in [0,1]$. For users who choose $a = F$, mean-variance preferences imply*

$$u_t^{MV}(P, q_t) = \bar{U} + R - c - Pq_t - \psi_t P^2 q_t(1-q_t),$$

where $\bar{U} \equiv U(s_{i,k})$ is deterministic and $\psi_t \geq 0$ is the type-specific risk aversion parameter.

Fix a reference mechanism $(\bar{P}, \bar{q}_t)$ with $\bar{P} \geq 0$ and $\bar{q}_t \in [0,1]$, and define

$$\phi_t \equiv 1 + \psi_t\, \bar{P}(1-\bar{q}_t).$$

Then for any (P, q_t),

$$u_t^{MV}(P, q_t) = \bar{U} + R - c - \phi_t\, Pq_t + \Delta_t(P, q_t),$$

where the approximation error is given by

$$\Delta_t(P, q_t) = -\psi_t P q_t \left(P(1-q_t) - \bar{P}(1-\bar{q}_t)\right).$$

Moreover, the error satisfies the bound

$$|\Delta_t(P, q_t)| \leq \psi_t\, Pq_t \left(|P-\bar{P}| + |q_t - \bar{q}_t|\, \max\{P, \bar{P}\}\right).$$

In particular, when (P, q_t) lies in a neighborhood of $(\bar{P}, \bar{q}_t)$, mean-variance utility is well approximated by the reduced form

$$u_t^{RF}(P, q_t) \equiv \bar{U} + R - c - \phi_t\, Pq_t,$$

which preserves the endogenous incentive term Pq_t while absorbing the second-order risk premium into a constant type-specific coefficient.

Proof Starting from mean-variance utility,

$$u_t^{MV}(P, q_t) = \bar{U} + R - c - Pq_t - \psi_t P^2 q_t(1-q_t) = \bar{U} + R - c - Pq_t[1 + \psi_t P(1-q_t)].$$

Add and subtract $\psi_t \bar{P}(1-\bar{q}_t)$ inside the bracket to obtain

$$u_t^{MV}(P, q_t) = \bar{U} + R - c - Pq_t[1 + \psi_t \bar{P}(1-\bar{q}_t)] - \psi_t P q_t[P(1-q_t) - \bar{P}(1-\bar{q}_t)].$$

Defining $\phi_t = 1 + \psi_t \bar{P}(1-\bar{q}_t)$ yields the stated decomposition, with

$$\Delta_t(P, q_t) = -\psi_t P q_t[P(1-q_t) - \bar{P}(1-\bar{q}_t)].$$

The bound follows from the triangle inequality and the facts that $P(1-q_t) \in [0, P]$ and $\bar{P}(1-\bar{q}_t) \in [0, \bar{P}]$. □

Fixing a reference mechanism $(\bar{P}, \bar{q}_t)$, Lemma 7 shows that mean-variance utility can be written as the reduced form

$$\bar{U} + R - c - \phi_t P q_t, \qquad \phi_t = 1 + \psi_t \bar{P}(1 - \bar{q}_t),$$

up to an approximation error that is small in a neighborhood of the reference point.

B.2 Proof of Theorem 1

We prove both necessity and sufficiency of the separation condition.

Necessity: Suppose perfect separation equilibrium exists with $p_H = 1$ and $p_L = 0$.

Since high-types participate with probability 1, their participation constraint must be satisfied:

$$R - c - \phi_H P \rho (1 - \eta_H) \geq 0 \tag{PC-H}$$

Since low-types abstain with probability 1, they must strictly prefer not participating:

$$R - c - \phi_L P \rho (1 - \eta_L) < 0 \tag{PC-L}$$

For the platform to minimize costs while maintaining separation, the high-type constraint binds:

$$R = c + \phi_H P \rho (1 - \eta_H) \tag{B1}$$

Substituting (B1) into (PC-L):

$$c + \phi_H P \rho (1 - \eta_H) - c - \phi_L P \rho (1 - \eta_L) < 0 \tag{B2}$$

$$P \rho [\phi_H (1 - \eta_H) - \phi_L (1 - \eta_L)] < 0 \tag{B3}$$

If $P = 0$ or $\rho = 0$, then both participation constraints become $R - c \geq 0$ and $R - c < 0$, which is impossible. Therefore, we must have $P > 0$ and $\rho > 0$ in any meaningful separation equilibrium.

With $P\rho > 0$, we obtain:

$$\phi_H (1 - \eta_H) < \phi_L (1 - \eta_L) \tag{B4}$$

Sufficiency: Suppose $\phi_H (1 - \eta_H) < \phi_L (1 - \eta_L)$ holds.

Set $R = c + \phi_H P \rho (1 - \eta_H)$ for some $P > 0$ and $\rho > 0$.

High-types' expected utility from participation is:

$$U_H = R - c - \phi_H P \rho (1 - \eta_H) = 0 \geq 0 \tag{B5}$$

So high-types are willing to participate ($p_H = 1$ is optimal).

Low-types' expected utility from participation would be:

$$U_L = R - c - \phi_L P\rho(1-\eta_L) \tag{B6}$$
$$= \phi_H P\rho(1-\eta_H) - \phi_L P\rho(1-\eta_L) \tag{B7}$$
$$= P\rho[\phi_H(1-\eta_H) - \phi_L(1-\eta_L)] < 0 \tag{B8}$$

So low-types strictly prefer to abstain ($p_L = 0$ is optimal).

Therefore, perfect separation equilibrium exists.

The equivalent ratio form follows by dividing both sides of the inequality by $\phi_L(1-\eta_H) > 0$:

$$\frac{\phi_H}{\phi_L} < \frac{1-\eta_L}{1-\eta_H} \tag{B9}$$

B.3 Proof of Theorem 2

Under perfect separation, we have $\theta^{F,H} = \lambda$, $\theta^{F,L} = 0$, $\theta^{v,H} = \rho\lambda$, $\theta^{v,L} = 0$, and $\theta^u = (1-\rho)\lambda$. The platform's profit function is

$$\Pi(\rho, R, P) = \gamma_v \eta_H \rho\lambda + \gamma_u(\eta_u^+ - \eta_u^-)(1-\rho)\lambda - \frac{k}{2}(\rho\lambda)^2 - R\lambda + P\rho(1-\eta_H)\lambda - \alpha P^2\lambda. \tag{B10}$$

In any perfect separation equilibrium, the high-type participation constraint binds, so

$$R = c + \phi_H P\rho(1-\eta_H). \tag{B11}$$

Substituting into Π yields the reduced objective

$$\Pi(\rho, P) = \lambda\Big[\gamma_v \eta_H \rho + \gamma_u(\eta_u^+ - \eta_u^-)(1-\rho) - \frac{k\lambda}{2}\rho^2 - c\Big] + \lambda P\rho(1-\eta_H)(1-\phi_H) - \lambda\alpha P^2. \tag{B12}$$

The choice variables satisfy $\rho \in [0,1]$ and $P \geq 0$.

Case 2(i) in Theorem 2: $\phi_H \geq 1$.

When $\phi_H \geq 1$, we have $(1-\phi_H) \leq 0$. Hence, for any fixed $\rho \geq 0$, the term

$$\lambda P\rho(1-\eta_H)(1-\phi_H) - \lambda\alpha P^2 \tag{B13}$$

is weakly decreasing in P on $[0,\infty)$, so the optimal penalty is

$$P^* = 0. \tag{B14}$$

With $P^* = 0$, maximizing (B12) over $\rho \in [0,1]$ gives

$$\left.\frac{\partial \Pi}{\partial \rho}\right|_{P=0} = \lambda\Big[\gamma_v \eta_H - \gamma_u(\eta_u^+ - \eta_u^-) - k\lambda\rho\Big], \tag{B15}$$

so the optimal verification rate is

$$\rho^* = \min\left\{1, \max\left\{0, \frac{\gamma_v \eta_H - \gamma_u(\eta_u^+ - \eta_u^-)}{k\lambda}\right\}\right\}. \tag{B16}$$

Finally, $R^* = c$ follows from the binding constraint $R = c + \phi_H P \rho(1-\eta_H)$ with $P^* = 0$. This establishes Case 2(i) of the theorem.

Case 1 and Case 2(ii) in Theorem 2: $\phi_H < 1$**.**

When $\phi_H < 1$, we have $(1-\phi_H) > 0$. For any fixed $\rho \in [0,1]$, the reduced profit (B12) is a strictly concave quadratic in P (since the coefficient on P^2 is $-\lambda\alpha < 0$). Therefore, for each ρ the unique maximizer over $P \geq 0$ satisfies the first-order condition

$$\frac{\partial \Pi}{\partial P} = \lambda\Big[\rho(1-\eta_H)(1-\phi_H) - 2\alpha P\Big] = 0, \tag{B17}$$

which yields

$$P^*(\rho) = \frac{\rho(1-\eta_H)(1-\phi_H)}{2\alpha}. \tag{B18}$$

Substituting (B18) back into (B12) gives a one-dimensional problem in ρ:

$$\Pi(\rho, P^*(\rho)) = \lambda\Big[\gamma_u(\eta_u^+ - \eta_u^-) - c + \rho\Big(\gamma_v \eta_H - \gamma_u(\eta_u^+ - \eta_u^-)\Big) - \frac{k\lambda}{2}\rho^2\Big] + \lambda \cdot \frac{\rho^2(1-\eta_H)^2(1-\phi_H)^2}{4\alpha}. \tag{B19}$$

Equivalently, up to a ρ-independent constant,

$$\Pi(\rho, P^*(\rho)) = \text{const} + \lambda\Big(\gamma_v \eta_H - \gamma_u(\eta_u^+ - \eta_u^-)\Big)\rho - \frac{\lambda}{2}\Big(k\lambda - \frac{(1-\eta_H)^2(1-\phi_H)^2}{2\alpha}\Big)\rho^2. \tag{B20}$$

Case 1 of the theorem (interior/concave region): Suppose $\phi_H < 1$ and condition (5) holds, i.e.,

$$k\lambda - \frac{(1-\eta_H)^2(1-\phi_H)^2}{2\alpha} > 0. \tag{B21}$$

Then (B20) is strictly concave in ρ, and the unconstrained maximizer solves

$$\frac{\partial}{\partial \rho}\Pi(\rho, P^*(\rho)) = \lambda\Big(\gamma_v \eta_H - \gamma_u(\eta_u^+ - \eta_u^-)\Big) - \lambda\Big(k\lambda - \frac{(1-\eta_H)^2(1-\phi_H)^2}{2\alpha}\Big)\rho = 0, \tag{B22}$$

so

$$\tilde{\rho} = \frac{\gamma_v \eta_H - \gamma_u(\eta_u^+ - \eta_u^-)}{k\lambda - \frac{(1-\eta_H)^2(1-\phi_H)^2}{2\alpha}}. \tag{B23}$$

Imposing $\rho \in [0,1]$ yields

$$\rho^* = \min\left\{1, \max\left\{0, \frac{\gamma_v \eta_H - \gamma_u(\eta_u^+ - \eta_u^-)}{k\lambda - \frac{(1-\eta_H)^2(1-\phi_H)^2}{2\alpha}}\right\}\right\}. \tag{B24}$$

Given ρ^*, the optimal penalty follows from (B18):

$$P^* = \frac{\rho^*(1-\eta_H)(1-\phi_H)}{2\alpha}, \tag{B25}$$

and the reward is pinned down by the binding participation constraint:

$$R^* = c + \phi_H P^* \rho^* (1-\eta_H). \tag{B26}$$

This proves Case 1 of the theorem.

Case 2(ii) of the theorem (non-concave region): Now suppose $\phi_H < 1$ and condition (5) is violated, i.e.,

$$k\lambda - \frac{(1-\eta_H)^2(1-\phi_H)^2}{2\alpha} \leq 0. \tag{B27}$$

Then (B20) is weakly convex (linear if equality holds and strictly convex if the inequality is strict) in ρ. Therefore, the maximizer over the compact set $[0,1]$ must lie at a boundary point, so

$$\rho^* \in \{0, 1\}. \tag{B28}$$

To select between $\rho^* = 0$ and $\rho^* = 1$, compare $\Pi(1, P^*(1))$ and $\Pi(0, P^*(0))$ using (B19):

$$\Pi(1, P^*(1)) - \Pi(0, P^*(0)) = \lambda\Big[\gamma_v \eta_H - \gamma_u(\eta_u^+ - \eta_u^-) - \frac{k\lambda}{2} + \frac{(1-\eta_H)^2(1-\phi_H)^2}{4\alpha}\Big]. \tag{B29}$$

Hence,

$$\rho^* = \begin{cases} 1, & \text{if } \gamma_v \eta_H - \gamma_u(\eta_u^+ - \eta_u^-) - \dfrac{k\lambda}{2} + \dfrac{(1-\eta_H)^2(1-\phi_H)^2}{4\alpha} \geq 0, \\ 0, & \text{otherwise.} \end{cases} \tag{B30}$$

Given ρ^*, the penalty is again pinned down by (B18):

$$P^* = \frac{\rho^*(1-\eta_H)(1-\phi_H)}{2\alpha}, \tag{B31}$$

(which implies $P^* = 0$ when $\rho^* = 0$ and $P^* > 0$ when $\rho^* = 1$), and the reward follows from binding participation:

$$R^* = c + \phi_H P^* \rho^* (1-\eta_H). \tag{B32}$$

This proves Case 2(ii) of the theorem and completes the proof.

B.4 Reverse Screening: Proof of Theorem 4

Proof Fix any mechanism (ρ, R, P). A type $i \in \{H, L\}$ participates if and only if its expected payoff from participation is nonnegative, i.e.,

$$R - c \;\geq\; \phi_i \, P \, \rho \, (1-\eta_i). \tag{B33}$$

Therefore, a *strict* reverse-screening outcome $(p_H, p_L) = (0, 1)$ requires

$$\text{(PC-L)} \quad R - c \;\geq\; \phi_L P \rho (1-\eta_L), \tag{B34}$$

$$\text{(NP-H)} \quad R - c \;<\; \phi_H P \rho (1-\eta_H). \tag{B35}$$

(We implicitly require $P\rho > 0$; otherwise the two right-hand sides coincide and it is impossible to have $p_L = 1$ while $p_H = 0$.)

(⇒) Necessity.

If a strict reverse-screening equilibrium exists, then there is an R satisfying (B34) and (B35). Combining them yields

$$\phi_L P \rho (1-\eta_L) \;\leq\; R - c \;<\; \phi_H P \rho (1-\eta_H). \tag{B36}$$

Since $P\rho > 0$, dividing through by $P\rho$ gives

$$\phi_L (1-\eta_L) \;<\; \phi_H (1-\eta_H), \tag{B37}$$

which is (18). The ratio form (19) follows by dividing both sides of (18) by $\phi_L(1-\eta_H) > 0$ (when $\phi_L > 0$; if $\phi_L = 0$, the strict inequality reduces to $\phi_H(1-\eta_H) > 0$).

(⇐) Sufficiency.

Assume (18) and $P\rho > 0$. Then the interval

$$\Big[c + \phi_L P \rho (1-\eta_L), \;\; c + \phi_H P \rho (1-\eta_H)\Big) \tag{B38}$$

is nonempty. Pick any R in this interval. By construction, (PC-L) holds and (NP-H) holds, so $(p_H, p_L) = (0, 1)$ is implemented. Hence a strict reverse-screening equilibrium exists.

Fix any (ρ, P) that admits strict reverse screening. If R satisfies (PC-L) with slack, i.e., $R > c + \phi_L P\rho(1-\eta_L)$, then lowering R to $\tilde{R} := c + \phi_L P\rho(1-\eta_L)$ preserves low-type participation and weakly decreases high-type incentives to participate, so $(p_H, p_L) = (0,1)$ is preserved. Since R enters the platform objective as a pure transfer when low types participate, an optimizing platform prefers the smallest feasible reward. Therefore, in any optimal reverse-screening mechanism,

$$R \;=\; c + \phi_L\, P\, \rho\, (1-\eta_L), \tag{B39}$$

which is (20). □

B.5 Proof of Theorem 5

Under reverse screening, low types participate and high types abstain, so

$$\theta^{F,L} = 1-\lambda, \qquad \theta^{F,H} = 0,$$

and therefore

$$\theta^{v,L} = \rho(1-\lambda), \qquad \theta^{v,H} = 0, \qquad \theta^{u} = (1-\rho)(1-\lambda).$$

The platform's profit function becomes

$$\begin{aligned}\Pi(\rho, R, P) = \gamma_v \eta_L \rho(1-\lambda) + \gamma_u(\eta_u^+ - \eta_u^-)(1-\rho)(1-\lambda) - \frac{k}{2}\big(\rho(1-\lambda)\big)^2 \\ - R(1-\lambda) + P\rho(1-\eta_L)(1-\lambda) - \alpha P^2(1-\lambda).\end{aligned} \tag{B40}$$

In any reverse-screening outcome with $(p_H, p_L) = (0,1)$, the low-type participation constraint binds:

$$R = c + \phi_L P\rho(1-\eta_L). \tag{B41}$$

Moreover, deterring high types requires

$$R \le c + \phi_H P\rho(1-\eta_H), \tag{B42}$$

which, using (B41), is equivalent to the reverse-screening condition $\phi_L(1-\eta_L) \le \phi_H(1-\eta_H)$ in (21) (strict inequality yields strict deterrence).

Substituting (B41) into (B40) yields the reduced objective

$$\begin{aligned}\Pi(\rho, P) = (1-\lambda)\Big[\gamma_v \eta_L \rho + \gamma_u(\eta_u^+ - \eta_u^-)(1-\rho) - \frac{k(1-\lambda)}{2}\rho^2 - c\Big] \\ + (1-\lambda)P\rho(1-\eta_L)(1-\phi_L) - (1-\lambda)\alpha P^2,\end{aligned} \tag{B43}$$

with $\rho \in [0,1]$ and $P \ge 0$.

Case 2(i) in Theorem 5: $\phi_L \geq 1$**.**

When $\phi_L \geq 1$, we have $(1-\phi_L) \leq 0$. Hence, for any fixed $\rho \geq 0$, the term $(1-\lambda)\big[P\rho(1-\eta_L)(1-\phi_L) - \alpha P^2\big]$ is weakly decreasing in P on $[0,\infty)$, so $P^* = 0$. With $P^* = 0$, maximizing (B43) over $\rho \in [0,1]$ yields

$$\rho^* = \min\left\{1, \max\left\{0, \frac{\gamma_v \eta_L - \gamma_u(\eta_u^+ - \eta_u^-)}{k(1-\lambda)}\right\}\right\}, \tag{B44}$$

and $R^* = c$ follows from (B41). This establishes Case 2(i).

Case 1 and Case 2(ii) in Theorem 5: $\phi_L < 1$**.**

When $\phi_L < 1$, we have $(1-\phi_L) > 0$. For any fixed $\rho \in [0,1]$, the reduced profit (B43) is strictly concave in P, so the unique maximizer satisfies

$$\frac{\partial \Pi}{\partial P} = (1-\lambda)\Big[\rho(1-\eta_L)(1-\phi_L) - 2\alpha P\Big] = 0, \tag{B45}$$

giving

$$P^*(\rho) = \frac{\rho(1-\eta_L)(1-\phi_L)}{2\alpha}. \tag{B46}$$

Substituting (B46) into (B43) yields, up to a ρ-independent constant,

$$\Pi(\rho, P^*(\rho)) = \text{const} + (1-\lambda)\Big(\gamma_v\eta_L - \gamma_u(\eta_u^+ - \eta_u^-)\Big)\rho - \frac{(1-\lambda)}{2}\Big(k(1-\lambda) - \frac{(1-\eta_L)^2(1-\phi_L)^2}{2\alpha}\Big)\rho^2. \tag{B47}$$

Case 1 (interior/concave region): If

$$k(1-\lambda) > \frac{(1-\eta_L)^2(1-\phi_L)^2}{2\alpha}, \tag{B48}$$

then (B47) is strictly concave in ρ, and the unconstrained maximizer is

$$\tilde{\rho} = \frac{\gamma_v\eta_L - \gamma_u(\eta_u^+ - \eta_u^-)}{k(1-\lambda) - \frac{(1-\eta_L)^2(1-\phi_L)^2}{2\alpha}}. \tag{B49}$$

Imposing $\rho \in [0,1]$ yields the stated ρ^*. Then the expressions for P^* and R^* follow from (B46) and (B41). This proves Case 1.

Case 2(ii) (non-concave region): If (B48) is violated, then (B47) is weakly convex in ρ, so the maximizer over $[0,1]$ satisfies $\rho^* \in \{0,1\}$. Comparing $\Pi(1, P^*(1))$ and

$\Pi(0, P^*(0))$ yields

$$\Pi(1, P^*(1)) - \Pi(0, P^*(0)) = (1-\lambda)\Big[\gamma_v \eta_L - \gamma_u(\eta_u^+ - \eta_u^-) - \frac{k(1-\lambda)}{2} + \frac{(1-\eta_L)^2(1-\phi_L)^2}{4\alpha}\Big], \tag{B50}$$

which implies the boundary rule for ρ^*. The expressions for P^* and R^* follow from (B46) and (B41). This completes the proof.

B.6 Proof of Proposition 6

Proof Work under perfect separation, so $\theta^{F,H} = \lambda$ and $\theta^{F,L} = 0$. Under the interior case with $\phi_H < 1$, Theorem 2 implies

$$\rho^*(\lambda) = \frac{N}{k\lambda - D}, \qquad P^*(\lambda) = \frac{\rho^*(\lambda)(1-\eta_H)(1-\phi_H)}{2\alpha}, \qquad R^*(\lambda) = c + \phi_H P^*(\lambda)\rho^*(\lambda)(1-\eta_H), \tag{B51}$$

where N and D are defined in (36).

Substituting the binding reward condition $R = c + \phi_H P \rho (1-\eta_H)$ into the platform profit under perfect separation yields, for any (ρ, P),

$$\Pi(\rho, P; \lambda) = \lambda\Big[\gamma_v \eta_H \rho + \gamma_u(\eta_u^+ - \eta_u^-)(1-\rho) - c\Big] - \frac{k}{2}(\rho\lambda)^2 + \lambda P \rho (1-\eta_H)(1-\phi_H) - \alpha P^2 \lambda. \tag{B52}$$

When $\phi_H < 1$, maximizing (B52) over $P \geq 0$ for a fixed ρ yields the unique best response

$$P^*(\rho) = \frac{\rho(1-\eta_H)(1-\phi_H)}{2\alpha}. \tag{B53}$$

Substituting (B53) into (B52) gives

$$\Pi(\rho, P^*(\rho); \lambda) = \lambda\big[\gamma_u(\eta_u^+ - \eta_u^-) - c\big] + \lambda N \rho - \frac{\lambda}{2}(k\lambda - D)\rho^2. \tag{B54}$$

The first-order condition in ρ from (B54) yields the interior optimizer

$$\rho^*(\lambda) = \frac{N}{k\lambda - D}, \tag{B55}$$

which is feasible exactly when (35) holds. Plugging (B55) back into (B54) gives optimal profit

$$\Pi^*(\lambda) = \lambda\big[\gamma_u(\eta_u^+ - \eta_u^-) - c\big] + \frac{\lambda N^2}{2(k\lambda - D)}. \tag{B56}$$

Differentiate (B56) with respect to λ:

$$\begin{aligned} \frac{d\Pi^*}{d\lambda} &= \big[\gamma_u(\eta_u^+ - \eta_u^-) - c\big] + \frac{N^2}{2(k\lambda - D)} - \frac{k\lambda N^2}{2(k\lambda - D)^2} \\ &= \big[\gamma_u(\eta_u^+ - \eta_u^-) - c\big] - \frac{DN^2}{2(k\lambda - D)^2}. \end{aligned} \tag{B57}$$

Since $D > 0$, the second term in (B57) is strictly negative. Moreover, as $\lambda \downarrow D/k$ (the lower boundary of the interior region), we have $(k\lambda - D)^2 \downarrow 0$ and thus $\frac{DN^2}{2(k\lambda - D)^2} \to +\infty$, implying $\frac{d\Pi^*}{d\lambda} \to -\infty$. Hence $\Pi^*(\lambda)$ is decreasing for λ sufficiently close to D/k from above.

To ensure the existence of some $\bar{\lambda} \in (0, 1)$ with negative slope while not having $\Pi^*(\lambda)$ decreasing everywhere, it suffices that the derivative is positive at $\lambda = 1$:

$$\left.\frac{d\Pi^*}{d\lambda}\right|_{\lambda=1} = \big[\gamma_u(\eta_u^+ - \eta_u^-) - c\big] - \frac{DN^2}{2(k-D)^2} \;>\; 0, \tag{B58}$$

which is equivalent to condition (37). Under (37), the derivative is positive at $\lambda = 1$ but negative for λ close enough to D/k, so by continuity there exists $\bar{\lambda} \in (D/k, 1)$ such that $\frac{d\Pi^*}{d\lambda}\big|_{\lambda=\bar{\lambda}} < 0$. This proves that $\Pi^*(\lambda)$ is not monotonically increasing on $(0, 1)$. □

References

[1] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., *et al.*: Training language models to follow instructions with human feedback. Advances in neural information processing systems **35**, 27730–27744 (2022)

[2] Schubert, L., Tong, M.: Extracting and evaluating general world knowledge from the brown corpus. In: Proceedings of the HLT-NAACL 2003 Workshop on Text Meaning, pp. 7–13 (2003)

[3] Townsend, R.M.: Optimal contracts and competitive markets with costly state verification. Journal of Economic theory **21**(2), 265–293 (1979)

[4] Gale, D., Hellwig, M.: Incentive-compatible debt contracts: The one-period problem. The Review of Economic Studies **52**(4), 647–663 (1985)

[5] Kofman, F., Lawarrée, J.: Collusion in hierarchical agency. Econometrica: Journal of the Econometric Society, 629–656 (1993)

[6] Duetting, P., Mirrokni, V., Paes Leme, R., Xu, H., Zuo, S.: Mechanism design for large language models. In: Proceedings of the ACM Web Conference 2024, pp. 144–155 (2024)

[7] Tang, Y., Feng, Y.: Beyond pairwise: Empowering llm alignment with ranked choice modeling. arXiv preprint arXiv:2510.23631 (2025)

[8] Fibich, G., Gavious, A., Sela, A.: All-pay auctions with risk-averse players. International Journal of Game Theory **34**(4), 583–599 (2006)

[9] Rochet, J.-C., Tirole, J.: Platform competition in two-sided markets. Journal of the european economic association **1**(4), 990–1029 (2003)

[10] Armstrong, M.: Competition in two-sided markets. The RAND journal of economics **37**(3), 668–691 (2006)

[11] Cabral, L., Hortacsu, A.: The dynamics of seller reputation: Evidence from ebay. The journal of industrial economics **58**(1), 54–78 (2010)

[12] Miller, N., Resnick, P., Zeckhauser, R.: Eliciting informative feedback: The peer-prediction method. Management Science **51**(9), 1359–1373 (2005)

[13] Prelec, D.: A bayesian truth serum for subjective data. science **306**(5695), 462–466 (2004)

[14] Dasgupta, A., Ghosh, A.: Crowdsourced judgement elicitation with endogenous proficiency. In: Proceedings of the 22nd International Conference on World Wide Web, pp. 319–330 (2013)

[15] Kong, Y., Schoenebeck, G.: Water from two rocks: Maximizing the mutual information. In: Proceedings of the 2018 ACM Conference on Economics and Computation, pp. 177–194 (2018)

[16] Northcutt, C., Jiang, L., Chuang, I.: Confident learning: Estimating uncertainty in dataset labels. Journal of Artificial Intelligence Research **70**, 1373–1411 (2021)

[17] Settles, B.: Active learning literature survey (2009)

[18] Zheng, Z., Peng, Y., Wu, F., Tang, S., Chen, G.: An online pricing mechanism for mobile crowdsensing data markets. In: Proceedings of the 18th ACM International Symposium on Mobile Ad Hoc Networking and Computing, pp. 1–10 (2017)

[19] Tadelis, S.: Reputation and feedback systems in online platform markets. Annual review of economics **8**(1), 321–340 (2016)

[20] Shah, N.B., Zhou, D.: Double or nothing: Multiplicative incentive mechanisms for crowdsourcing. Journal of Machine Learning Research **17**(165), 1–52 (2016)

[21] Sun, H., Chen, Y., Wang, S., Chen, W., Deng, X.: Mechanism design for llm fine-tuning with multiple reward models. arXiv preprint arXiv:2405.16276 (2024)

[22] Liu, Y., Chen, Y.: Machine-learning aided peer prediction. In: Proceedings of the 2017 ACM Conference on Economics and Computation, pp. 63–80 (2017)