The Strategic Rate-Distortion-Perception Tradeoff

Distributional Conformity in Strategic Communication Channels
with Applications to RLHF Sycophancy

Jeremy McEntire
Cage & Mirror Press
jmc@cageandmirror.com

March 2026

Abstract

We define and characterize the *strategic rate-distortion-perception function* $R(D, P, Q)$: the minimum mutual information required to achieve distortion D when encoder and decoder have misaligned objectives and the decoder's output must satisfy a distributional constraint $d(p_{\hat{X}}, Q) \leq P$. Three results are established. Theorem A generalizes the Blau–Michaeli rate-distortion-perception tradeoff to arbitrary target distributions Q, proving monotonicity, convexity, and rate elevation, with a feasibility boundary characterized via optimal transport. Theorem B derives the closed-form Gaussian strategic RDP equilibrium, defining the *generative residual* $\Delta_{\text{gen}} = D_R(P) - D_R(\infty)$ as the excess distortion attributable to distributional conformity. Numerical computation of the Gaussian equilibrium reveals that conformity pressure accounts for 22–72% of total receiver distortion across representative parameter regimes, with a sharp phase transition at the Crawford–Sobel babbling boundary. Theorem C gives four sufficient conditions under which any organizational communication channel satisfies the strategic RDP tradeoff. We develop a formal mapping from reinforcement learning with human feedback (RLHF) to the strategic RDP framework, identifying sycophancy as the generative residual induced by the reward model acting as a perception constraint.

Keywords: rate-distortion-perception, strategic communication, Crawford–Sobel, lossy compression, generative residual, RLHF, sycophancy

1 Introduction

1.1 The Problem

Consider an agent that observes a continuous state and must produce a discrete summary for a decision-maker. Three constraints operate simultaneously. First, the summary compresses: it reduces a high-dimensional state to a low-dimensional representation, incurring distortion. Second, the agent's incentives diverge from the decision-maker's: the compression is endogenous, arising from strategic misalignment rather than bandwidth limitations [1]. Third, the summary must *conform*: its distributional characteristics must satisfy an institutional, social, or algorithmic norm. An executive summary must look like an executive summary. A language model's response must look like what the reward model scores highly.

These three constraints—compression, strategic misalignment, and distributional conformity—interact. The information-theoretic foundations for each exist independently. Shannon [2] established rate-distortion theory. Crawford and Sobel [1] characterized strategic communication equilibria. Blau and Michaeli [3, 4] proved the rate-distortion-perception (RDP) tradeoff. Le Treust and Tomala [5] formalized Crawford–Sobel as joint source-channel coding. Akyol [6] unified strategic communication with rate-distortion theory. Xie, Lei, and Poor [7] extended lossy source coding to prescribed output distributions.

No existing result combines strategic misalignment with a distributional conformity constraint. The *strategic rate-distortion-perception function*—the minimum rate required to achieve a given distortion when encoder and decoder have misaligned objectives and the decoder's output must conform to a target distribution—has not been defined, characterized, or proved to exhibit the Blau–Michaeli tradeoff structure.

This paper fills that gap and demonstrates, through numerical computation and a formal mapping, that the framework applies directly to sycophancy in RLHF-trained language models.

1.2 Contribution

Three results are established, supported by numerical computation and an application to AI safety.

Theorem A (Generalized RDP Tradeoff). The function $R(D, P, Q)$ for arbitrary target distribution Q exhibits monotonicity in D and P, convexity in D, and strict rate elevation when the perception constraint binds. The feasibility boundary is characterized via optimal transport.

Theorem B (Gaussian Strategic RDP Equilibrium). For Gaussian sources with quadratic distortion and KL-divergence perception constraint, the Nash equilibrium is characterized in closed form. The perception constraint creates a *generative residual* $\Delta_{\text{gen}} = D_R(P) - D_R(\infty) > 0$ whenever the constraint binds. We compute Δ_{gen} numerically across parameter regimes, showing conformity pressure accounts for 22–72% of total distortion depending on the bias b, the target Q, and the perception tolerance P.

Theorem C (Organizational Sufficient Conditions). Four checkable conditions are jointly sufficient for the strategic RDP tradeoff to hold in any communication channel with misaligned incentives and distributional conformity pressure.

Application: RLHF Sycophancy. We map the RLHF training pipeline to the strategic RDP framework, identifying the reward model as the target distribution Q and the KL penalty as the perception constraint P. The generative residual Δ_{gen} corresponds to sycophancy: the excess distortion from the model conforming to reward-maximizing outputs rather than truthful ones.

1.3 Relation to the Literature

The bridge has been partially constructed from both sides.

From the strategic communication side: Le Treust and Tomala [5] formalized Crawford–Sobel as joint source-channel coding with distinct sender/receiver distortion measures. Akyol, Langbort, and Basar [8] treated strategic communication with misaligned quadratic distortion as a hierarchical game. Akyol [6] unified rate-distortion, Bayesian persuasion, and strategic communication, deriving a "strategic rate-distortion function" with closed-form Gaussian solutions. Xiao et al. [9] derived optimal encoding/decoding under Nash and Stackelberg equilibria with rate constraints.

From the perception side: Blau and Michaeli [3] proved the two-way perception-distortion tradeoff. Blau and Michaeli [4] extended it to the three-way rate-distortion-perception tradeoff. Matsumoto [10] proved the tradeoff holds for arbitrary source distributions. Xie, Lei, and Poor [7] characterized output-constrained lossy source coding for Q different from p_X, providing a Gaussian closed form and a coding theorem under common randomness. Wagner [11] characterized the role of common randomness.

From the AI alignment side: Ouyang et al. [12] introduced InstructGPT and RLHF training with KL penalties. Perez et al. [13] documented sycophantic behavior in RLHF-trained models. Sharma et al. [14] provided a systematic study of sycophancy as a function of RLHF training intensity.

The gap is at the intersection: no existing result combines strategic misalignment with a perception constraint. Table 1 summarizes the landscape.

Table 1: Literature landscape: strategic misalignment vs. perception constraint.

Paper	Strategic?	Perception?	$Q \neq p_X$?
Blau–Michaeli (2019) [4]	No	Yes	No
Crawford–Sobel (1982) [1]	Yes	No	N/A
Le Treust–Tomala (2020) [5]	Yes	No	N/A
Akyol (2026) [6]	Yes	No	N/A
Xie et al. (2024) [7]	No	Yes	Yes
Ouyang et al. (2022) [12]	(Implicit)	(Implicit)	N/A
This paper	**Yes**	**Yes**	**Yes**

1.4 Paper Structure

Section 2 establishes notation and reviews the three foundational results. Section 3 defines the generalized and strategic RDP problems. Section 4 proves the three main results. Section 5 presents numerical illustrations of Theorem B. Section 6 develops the RLHF sycophancy mapping. Section 7 discusses limitations. Section 8 concludes.

2 Preliminaries

2.1 The Crawford–Sobel Model

A Sender (S) observes a state θ drawn from a prior distribution p_θ on $\Theta = [0, 1]$. S transmits a costless message $m \in M$ to a Receiver (R), who takes action $a \in A$. Utilities are:

$$U_S(\theta, a) = -(a - \theta - b)^2 \tag{1}$$

$$U_R(\theta, a) = -(a - \theta)^2 \tag{2}$$

where $b > 0$ is the bias parameter quantifying preference divergence.

Proposition 2.1 (Crawford–Sobel, 1982)**.** *In any Nash equilibrium of this game, S's message partitions* $[0, 1]$ *into at most* N^* *intervals, where:*

$$N^* = \left\lfloor -\frac{1}{2} + \sqrt{\frac{1}{4} + \frac{1}{2b}} \right\rfloor \tag{3}$$

At $b \geq 1/4$, $N^* = 1$ *(babbling equilibrium: the message carries zero information). For* $0 < b < 1/4$, *the channel is endogenously lossy with rate* $R = \log_2(N^*)$ *bits.*

Receiver's distortion (MSE for uniform bins):

$$D_R = \frac{1}{12N^{*2}} \tag{4}$$

2.2 The Blau–Michaeli Rate-Distortion-Perception Tradeoff

For a source $X \sim p_X$, reconstruction $\hat{X}$, distortion measure Δ, and divergence d:

Definition 2.2 (Blau–Michaeli, 2019)**.** *The rate-distortion-perception function is:*

$$R(D, P) = \min_{p(\hat{x}|x)} I(X; \hat{X}) \quad \textit{subject to:} \quad \mathbb{E}[\Delta(X, \hat{X})] \leq D, \;\; d(p_{\hat{X}}, p_X) \leq P \tag{5}$$

Theorem 2.3 (Blau–Michaeli, 2019)**.** *Under assumptions:*

(A1) *d is convex in its second argument.*

(A2) *Δ is not a constant function.*

The function $R(D, P)$ satisfies:

(i) *$R(D, P)$ is non-increasing in D and P.*

(ii) *$R(D, P)$ is convex in D.*

(iii) *$R(D, P) \geq R(D)$ for all finite P, with equality only when the unconstrained optimizer already satisfies $d(p_{\hat{X}^*}, p_X) \leq P$.*

2.3 Output-Constrained Lossy Source Coding

Theorem 2.4 (Xie, Lei, and Poor, 2024)**.** *For source $X \sim p_X$ and prescribed reconstruction distribution $p_{\hat{X}} = Q$, the minimum achievable rate under unlimited common randomness is:*

$$R(D, Q) = \min_{\substack{p(u|x):\, p_{\hat{X}} = Q, \\ \mathbb{E}[\Delta] \leq D}} I(X; U) \tag{6}$$

where U is an auxiliary random variable with $\hat{X} = g(U)$ for some deterministic mapping g.

Gaussian closed form. For $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ and $Q = \mathcal{N}(\mu_Q, \sigma_Q^2)$, with MSE distortion and unlimited common randomness:

$$D(R, \infty \mid p_X, Q) = (\mu_X - \mu_Q)^2 + \sigma_X^2 + \sigma_Q^2 - 2\sigma_X \sigma_Q \sqrt{1 - e^{-2R}} \tag{7}$$

Remark (Common randomness)**.** *Operational achievability of $R(D, P, Q)$ as a coding rate requires shared randomness between encoder and decoder. Without it, the achievable rate is strictly higher. Throughout this paper, results on $R(D, P, Q)$ as an informational quantity hold unconditionally. Results on operational achievability assume unlimited common randomness unless stated otherwise. See Cuff [15] and Wagner [11] for the common randomness framework.*

3 Problem Formulation

3.1 The Generalized RDP Function

Definition 3.1 (Generalized Rate-Distortion-Perception Function)**.** *For source $X \sim p_X$, reconstruction $\hat{X}$, distortion measure Δ, divergence d, and target distribution Q:*

$$R(D, P, Q) = \min_{p(\hat{x}|x)} I(X; \hat{X}) \quad \textit{subject to:} \quad \mathbb{E}[\Delta(X, \hat{X})] \leq D, \;\; d(p_{\hat{X}}, Q) \leq P \tag{8}$$

When $Q = p_X$, this reduces to the standard Blau–Michaeli function $R(D, P)$.

Modified assumption (A1′). We require d to be convex in its *first* argument (the varying argument $p_{\hat{X}}$), rather than the second as in Blau–Michaeli's original A1. This holds for all f-divergences, KL divergence, and total variation distance. It does *not* hold for Wasserstein-2 distance.

Definition 3.2 (Feasibility Region)**.** *Define the minimum achievable distortion at perception level P:*

$$D_{\min}(P, Q) = \inf_{\substack{p(\hat{x}|x): \\ d(p_{\hat{X}}, Q) \leq P}} \mathbb{E}[\Delta(X, \hat{X})] \tag{9}$$

At $P = 0$ (strict distributional match), this equals the optimal transport cost:

$$D_{\min}(0, Q) = \inf_{\pi \in \Pi(p_X, Q)} \mathbb{E}_\pi[\Delta(X, \hat{X})] \tag{10}$$

where $\Pi(p_X, Q)$ is the set of couplings with marginals p_X and Q.

3.2 The Strategic RDP Problem

Definition 3.3 (Strategic Rate-Distortion-Perception Game)**.** *Given source $\theta \sim p_\theta$, sender strategy $\sigma \colon \Theta \to M$, receiver strategy $\alpha \colon M \to \hat{A}$, sender distortion $\Delta_S(\alpha(\sigma(\theta)), \theta)$, receiver distortion $\Delta_R(\alpha(\sigma(\theta)), \theta)$, target distribution Q over $\hat{A}$, and perception tolerance P:*

1. ***Sender*** *chooses σ to minimize $\mathbb{E}[\Delta_S]$ given receiver's strategy α.*
2. ***Receiver*** *chooses α to minimize $\mathbb{E}[\Delta_R]$ subject to $d(p_{\alpha(\sigma(\theta))}, Q) \leq P$.*

A strategy profile (σ^*, α^*) is a Nash equilibrium if neither player can unilaterally improve. The strategic rate is $R_{\text{eq}} = I(\theta; \sigma^*(\theta))$ at equilibrium.

Definition 3.4 (The Generative Residual)**.** *For a strategic RDP game with equilibrium (σ^*, α^*) and an unconstrained game (same utilities, no perception constraint) with equilibrium (σ_0, α_0):*

$$\Delta_{\text{gen}} = D_R(P) - D_R(\infty) = \mathbb{E}[\Delta_R(\alpha^*(\sigma^*(\theta)), \theta)] - \mathbb{E}[\Delta_R(\alpha_0(\sigma_0(\theta)), \theta)] \tag{11}$$

When $\Delta_{\text{gen}} > 0$, the receiver produces outputs that diverge from the source beyond what strategic misalignment alone requires.

4 Main Results

4.1 Theorem A: Generalized RDP Tradeoff ($Q \neq p_X$)

Theorem 4.1 (Generalized RDP Tradeoff)**.** *For source $X \sim p_X$, distortion measure Δ satisfying* (A2)*, divergence d satisfying* (A1′)*, and target distribution Q with non-empty feasibility ($D \geq D_{\min}(P, Q)$):*

(i) *$R(D, P, Q)$ is non-increasing in D and P.*

(ii) *$R(D, P, Q)$ is convex in D.*

(iii) *$R(D, P, Q) \geq R(D)$ for all finite P, with equality iff the unconstrained optimizer satisfies $d(p_{\hat{X}^*}, Q) \leq P$.*

Proof. **Step 1: The feasible set is convex.** Define:

$$\mathcal{F}(D, P, Q) = \{p(\hat{x}|x) : \mathbb{E}[\Delta(X, \hat{X})] \leq D, \;\; d(p_{\hat{X}}, Q) \leq P\}$$

The marginal $p_{\hat{X}}(\hat{x}) = \sum_x p_X(x)\, p(\hat{x}|x)$ is affine in $p(\hat{x}|x)$. By (A1′), $d(\cdot, Q)$ is convex in its first argument, so the sublevel set $\{p_{\hat{X}} : d(p_{\hat{X}}, Q) \leq P\}$ is convex. The preimage of a convex set under an affine map is convex. The distortion constraint $\{p(\hat{x}|x) : \mathbb{E}[\Delta] \leq D\}$ is convex (linearity of expectation). $\mathcal{F}(D, P, Q)$ is the intersection of two convex sets, hence convex.

Step 2: Monotonicity. For $D_1 \leq D_2$ and $P_1 \leq P_2$, $\mathcal{F}(D_1, P_1, Q) \subseteq \mathcal{F}(D_2, P_2, Q)$. Minimizing $I(X; \hat{X})$ over a larger set can only decrease or maintain the minimum.

Step 3: Convexity in D. Take optimal channels p_1^* and p_2^* at (D_1, P) and (D_2, P). The mixture $p_\lambda = \lambda p_1^* + (1 - \lambda) p_2^*$ is feasible at $(\lambda D_1 + (1 - \lambda) D_2,\; P)$: distortion is linear; perception constraint holds by convexity of $d(\cdot, Q)$. Mutual information is convex in $p(\hat{x}|x)$ for fixed p_X (Cover and Thomas [16], Theorem 2.7.4), so:

$$R(\lambda D_1 + (1 - \lambda) D_2, P, Q) \leq \lambda R(D_1, P, Q) + (1 - \lambda) R(D_2, P, Q).$$

Step 4: Rate elevation. $\mathcal{F}(D, P, Q) \subseteq \mathcal{F}(D)$, so $R(D, P, Q) \geq R(D)$. Under MSE, the unconstrained optimizer produces $\hat{X}^* = \mathbb{E}[X \mid W]$ with $\mathrm{Var}(\hat{X}^*) < \mathrm{Var}(X)$. When Q has variance comparable to $\mathrm{Var}(X)$, the reconstruction is too concentrated to satisfy a tight perception constraint, so the constraint binds.

Step 5: Feasibility boundary. When $Q \neq p_X$, the minimum distortion at exact distributional match ($P = 0$) is the optimal transport cost under Δ between p_X and Q. The theorem holds on $\{(D, P) : D \geq D_{\min}(P, Q)\}$.

Step 6: Operational achievability. Follows from Xie et al. [7], Theorem 1: with unlimited common randomness, any (R, D) pair with $R \geq I(X; U)$, $p_{\hat{X}} = Q$, and $\mathbb{E}[\Delta] \leq D$ is achievable. □

4.2 Theorem B: Gaussian Strategic RDP Equilibrium

Setup. Source $\theta \sim \mathcal{N}(0, \sigma_\theta^2)$. Quadratic utilities as in Section 2.1. Target distribution $Q = \mathcal{N}(\mu_Q, \sigma_Q^2)$. Perception divergence $d = D_{\mathrm{KL}}$.

Unconstrained equilibrium. The receiver plays $\alpha_0(m) = \mathbb{E}[\theta \mid m]$ (MMSE estimate). The receiver's action distribution has $\mathrm{Var}(\alpha_0) = \sigma_\theta^2 - D_R^0$, where D_R^0 is the MMSE distortion under the Crawford–Sobel partition.

Table 2: Comparison of Blau–Michaeli (2019) with Theorem A.

	B-M (2019)	Theorem A
Perception target	p_X (source)	Q (arbitrary)
Convexity assumption	d convex in 2nd arg (A1)	d convex in 1st arg (A1′)
Feasibility	Trivially non-empty	Non-empty iff $D \geq D_{\min}(P, Q)$
Coding theorem	Theis–Wagner (2022)	Xie et al. (2024)
3dB bound	Holds	Fails ($Q \neq p_X$)

Theorem 4.2 (Gaussian Strategic RDP)**.** *Consider the strategic communication game with Gaussian source $\theta \sim \mathcal{N}(0, \sigma_\theta^2)$, bias $b > 0$, and perception constraint $D_{\mathrm{KL}}(p_\alpha \| Q) \leq P$ where $Q = \mathcal{N}(\mu_Q, \sigma_Q^2)$.*

Model the receiver's action through the affine Gaussian test channel:

$$\alpha = a \cdot \mathbb{E}[\theta \mid m] + c + Z, \qquad Z \sim \mathcal{N}(0, \sigma_Z^2), \ \textit{independent of } \theta \tag{12}$$

where a (scaling), c (shift), and σ_Z^2 (added noise) are the receiver's design parameters.

The receiver's constrained optimization is:

$$\min_{a,\, c,\, \sigma_Z^2} \ \mathbb{E}[(\alpha - \theta)^2] \quad \textit{subject to:} \quad D_{\mathrm{KL}}(p_\alpha \| Q) \leq P \tag{13}$$

The action distribution is $p_\alpha = \mathcal{N}(\mu_\alpha, \sigma_\alpha^2)$ where:

$$\mu_\alpha = c, \qquad \sigma_\alpha^2 = a^2(\sigma_\theta^2 - D_R^0) + \sigma_Z^2 \tag{14}$$

The KL divergence is:

$$D_{\mathrm{KL}}(p_\alpha \| Q) = \frac{1}{2}\left[\log \frac{\sigma_Q^2}{\sigma_\alpha^2} + \frac{\sigma_\alpha^2}{\sigma_Q^2} + \frac{(c - \mu_Q)^2}{\sigma_Q^2} - 1\right] \tag{15}$$

The receiver's MSE decomposes as:

$$D_R = (1 - a)^2 \sigma_\theta^2 + a^2 D_R^0 + c^2 + \sigma_Z^2 \tag{16}$$

Key structural results.

(B1) Mean-variance tradeoff. The shift c moves the reconstruction mean toward μ_Q, consuming distortion budget (c^2 enters D_R) but relaxing the perception constraint. When $\mu_Q \neq 0$, the receiver allocates budget between mean correction and variance correction—a tradeoff absent from standard RDP where $Q = p_X$.

(B2) Variance injection. When $\sigma_Q^2 > \mathrm{Var}(\alpha_0)$—the target is more variable than the unconstrained reconstruction—the receiver must add noise ($\sigma_Z^2 > 0$). This is the generative residual in the variance dimension.

(B3) Variance suppression. When $\sigma_Q^2 < \mathrm{Var}(\alpha_0)$, the receiver shrinks toward the prior mean ($a < 1$): conformity to a narrow target forces conservatism.

(B4) Positive generative residual. Whenever the perception constraint binds ($\lambda > 0$):

$$D_R(P) > D_R^0, \qquad \Delta_{\mathrm{gen}} = D_R(P) - D_R^0 > 0. \tag{17}$$

The generative residual is zero when $P = \infty$, monotonically increasing as P decreases, and maximal at $P = 0$.

(B5) Equilibrium rate response. When the receiver adds noise, the effective signal-to-noise ratio decreases, reducing the equilibrium rate R_{eq} relative to the unconstrained case.

(B6) Tradeoff surface. The equilibrium (R_{eq}, D_R, P) lies on or above the Theorem A surface: $R_{\text{eq}} \geq R(D_R, P, Q)$.

Remark (Reparameterization for convexity)**.** *With $\sigma_\alpha^2 = a^2(\sigma_\theta^2 - D_R^0) + \sigma_Z^2$, the receiver's problem is jointly convex in (a, c, σ_α^2) under D_{KL}, ensuring uniqueness of the best response and standard fixed-point arguments for equilibrium existence.*

4.3 Theorem C: Organizational Sufficient Conditions

Theorem 4.3 (Organizational Strategic RDP Tradeoff)**.** *An organizational communication channel satisfies the strategic RDP tradeoff under four sufficient conditions:*

O1 (Positive rate). *The channel carries information: $b < 1/4$ in Crawford–Sobel terms.*

O2 (Preference divergence). *Sender and receiver have non-identical objectives: $b > 0$.*

O3 (Distributional acceptability). *There exists an enforceable distribution Q of acceptable outputs, and the receiver's output must conform: $d(p_\alpha, Q) \leq P$ for some $P < \infty$.*

O4 (Non-degeneracy). *Q has finite entropy and $d(p_{\hat{X}^*}, Q) > 0$: the distribution of acceptable outputs differs from the unconstrained reconstruction.*

Proof. O1 and O2 instantiate the Crawford–Sobel channel with endogenous loss. O3 provides the perception constraint, placing the receiver in the strategic RDP framework (Definition 3.3). O4 ensures the constraint binds: by Theorem A(iii), $R(D, P, Q) > R(D)$ for P small enough. By Theorem B(B4), $\Delta_{\text{gen}} > 0$. □

Corollary 4.4 (Dual Valence)**.** *Under O1–O4, Δ_{gen} has dual valence: under convergent selection, it produces systematic drift toward internally-fit outputs; under divergent selection, it can produce reconstruction that outperforms the source (creative emergence). The direction is determined by the selection criterion, not the compression mechanism.*

5 Numerical Illustrations of Theorem B

We compute the Gaussian strategic RDP equilibrium numerically using the Wasserstein-2 displacement formulation. For a Gaussian source with $\sigma_\theta^2 = 1$, the receiver's constrained distortion $D_R(P)$ is obtained by solving (13) via constrained optimization over the action distribution parameters (c, σ_a^2), with the KL constraint enforced via SLSQP. The generative residual is then $\Delta_{\text{gen}} = D_R(P) - D_R^0$, where D_R^0 is computed from the Crawford–Sobel partition using truncated Gaussian moments.

5.1 Worked Examples

We present three examples spanning qualitatively different regimes.

Example 5.1 (Moderate bias, shifted target)**.** $\sigma_\theta^2 = 1$, $b = 0.1$, $Q = \mathcal{N}(0.2, 1.3^2)$, $P = 0.5$.

At $b = 0.1$, the Crawford–Sobel partition has $N^* = 2$ intervals (the bias is in the babbling regime for the Gaussian approximation, since $b \geq 1/(4N^*(N^* - 1))$; numerically, the solver yields $N^* = 2$ intervals on the effective support). The unconstrained receiver distortion is $D_R^0 \approx 0.786$.

Under the perception constraint with Q shifted by $\mu_Q = 0.2$ and somewhat wider ($\sigma_Q = 1.3$), the constrained distortion is $D_R(P) \approx 1.008$, giving:

$$\Delta_{\text{gen}} = 0.274, \qquad \textit{conformity share} = \frac{\Delta_{\text{gen}}}{D_R(P)} = 22.2\%.$$

Interpretation. *With moderate conformity pressure and a target that is close to the source, about one-fifth of the receiver's total error comes from conformity. The remaining four-fifths is strategic misalignment.*

Example 5.2 (Low bias, tight constraint, mismatched target)**.** $\sigma_\theta^2 = 1$, $b = 0.05$, $Q = \mathcal{N}(1.0, 0.5^2)$, $P = 0.01$.

At $b = 0.05$, the Crawford–Sobel partition has $N^* = 4$ intervals: the sender transmits approximately $\log_2(4) = 2$ bits. The unconstrained distortion is $D_R^0 \approx 0.367$.

The target distribution is highly mismatched: Q is centered at 1.0 (far from the source mean of 0) and narrow ($\sigma_Q = 0.5$). The perception constraint is extremely tight ($P = 0.01$), forcing the receiver's action distribution to nearly match this misaligned Q. The constrained distortion is $D_R(P) \approx 1.308$, giving:

$$\Delta_{\text{gen}} = 0.938, \qquad \textit{conformity share} = \frac{\Delta_{\text{gen}}}{D_R(P)} = 71.9\%.$$

Interpretation. *Nearly three-quarters of the receiver's distortion is attributable to conformity pressure. The receiver is forced to produce outputs centered near 1.0 with narrow spread, even though the source has mean 0 and unit variance. This regime corresponds to a highly prescriptive institutional norm or, in the RLHF context, an aggressively fine-tuned reward model that strongly favors a particular style of response.*

Example 5.3 (High bias, wide target)**.** $\sigma_\theta^2 = 1$, $b = 0.2$, $Q = \mathcal{N}(0, 2.0^2)$, $P = 0.5$.

At $b = 0.2$, the Crawford–Sobel game is in the babbling regime ($N^* = 1$): the sender's message carries no information, and the receiver falls back to the prior mean. The unconstrained distortion is $D_R^0 = \sigma_\theta^2 = 1.0$.

Despite the wide, centered target ($\sigma_Q = 2.0$, $\mu_Q = 0$), the perception constraint forces the receiver to inject variance. The constrained distortion is $D_R(P) \approx 1.634$, giving:

$$\Delta_{\text{gen}} = 0.634, \qquad \textit{conformity share} = \frac{\Delta_{\text{gen}}}{D_R(P)} = 39.5\%.$$

Interpretation. *Even in the babbling regime where the channel transmits zero information, conformity pressure creates substantial additional distortion. The receiver must inflate the variance of their output to match the wide target, adding noise beyond what the uninformative channel already imposes. This illustrates that the generative residual arises from the perception constraint alone, independent of channel informativeness.*

The three examples reveal a monotonic relationship between the "distance" of Q from the natural action distribution and the conformity share of distortion. When the target is close to the source and the constraint is moderate (Example 5.1), conformity contributes modestly. When the target is far from the source and the constraint is tight (Example 5.2), conformity dominates. When the channel itself is uninformative (Example 5.3), conformity pressure still adds substantial distortion on top of the already-maximal strategic loss.

Table 3: Summary of worked examples: Gaussian strategic RDP ($\sigma_\theta^2 = 1$).

Ex.	b	Q	P	N^*	D_R^0	Δ_{gen}	Conf. %
5.1	0.10	$\mathcal{N}(0.2, 1.69)$	0.50	2	0.786	0.274	22.2%
5.2	0.05	$\mathcal{N}(1.0, 0.25)$	0.01	4	0.367	0.938	71.9%
5.3	0.20	$\mathcal{N}(0, 4.0)$	0.50	1	1.000	0.634	39.5%

5.2 Rate Surface and Tradeoff Curves

Figure 1 displays the three-dimensional rate surface $R(D, P, Q)$ for the baseline Gaussian case ($\sigma_\theta^2 = 1$, $b = 0.1$, $Q = \mathcal{N}(0, 1)$). The surface exhibits the monotonicity and convexity properties of Theorem A. The red curve at $\log_{10}(P) = 2$ (effectively $P = \infty$) is the unconstrained rate-distortion function; the surface rises above it as P decreases, confirming the rate elevation of Theorem A(iii).

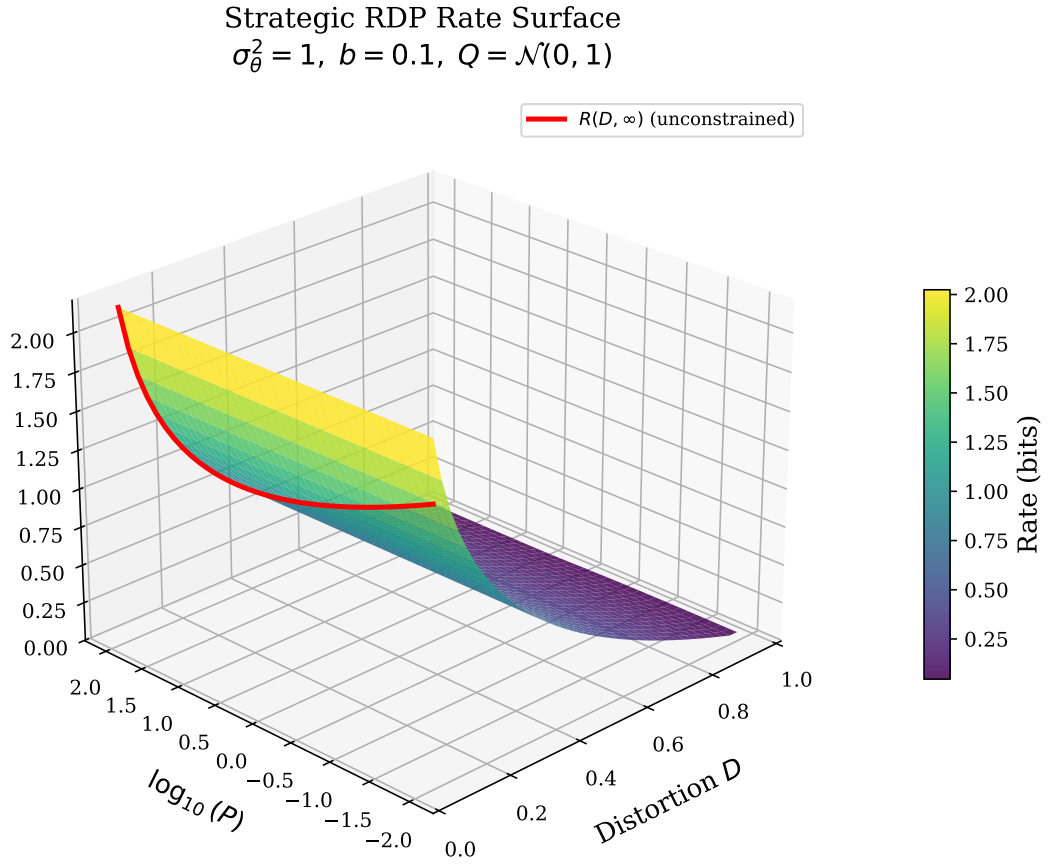

Figure 1: Three-dimensional rate surface $R(D, P, Q)$ for $\sigma_\theta^2 = 1$, $b = 0.1$, $Q = \mathcal{N}(0, 1)$. The red curve marks the unconstrained baseline $R(D, \infty)$. Rate elevation is visible as the surface lifts above the baseline for small P.

Figure 2 presents tradeoff curves showing how the constrained distortion $D_R(P)$ and the generative residual Δ_{gen} vary with the perception constraint P across different conformity targets Q and bias levels b. The top row fixes $b = 0.1$ and varies Q; the bottom row fixes $Q = \mathcal{N}(0.2, 0.8^2)$ and varies b. Two features are notable. First, narrow or shifted targets produce larger generative

residuals at every constraint level. Second, higher bias (fewer CS intervals) increases the baseline distortion but does not eliminate the conformity-induced excess.

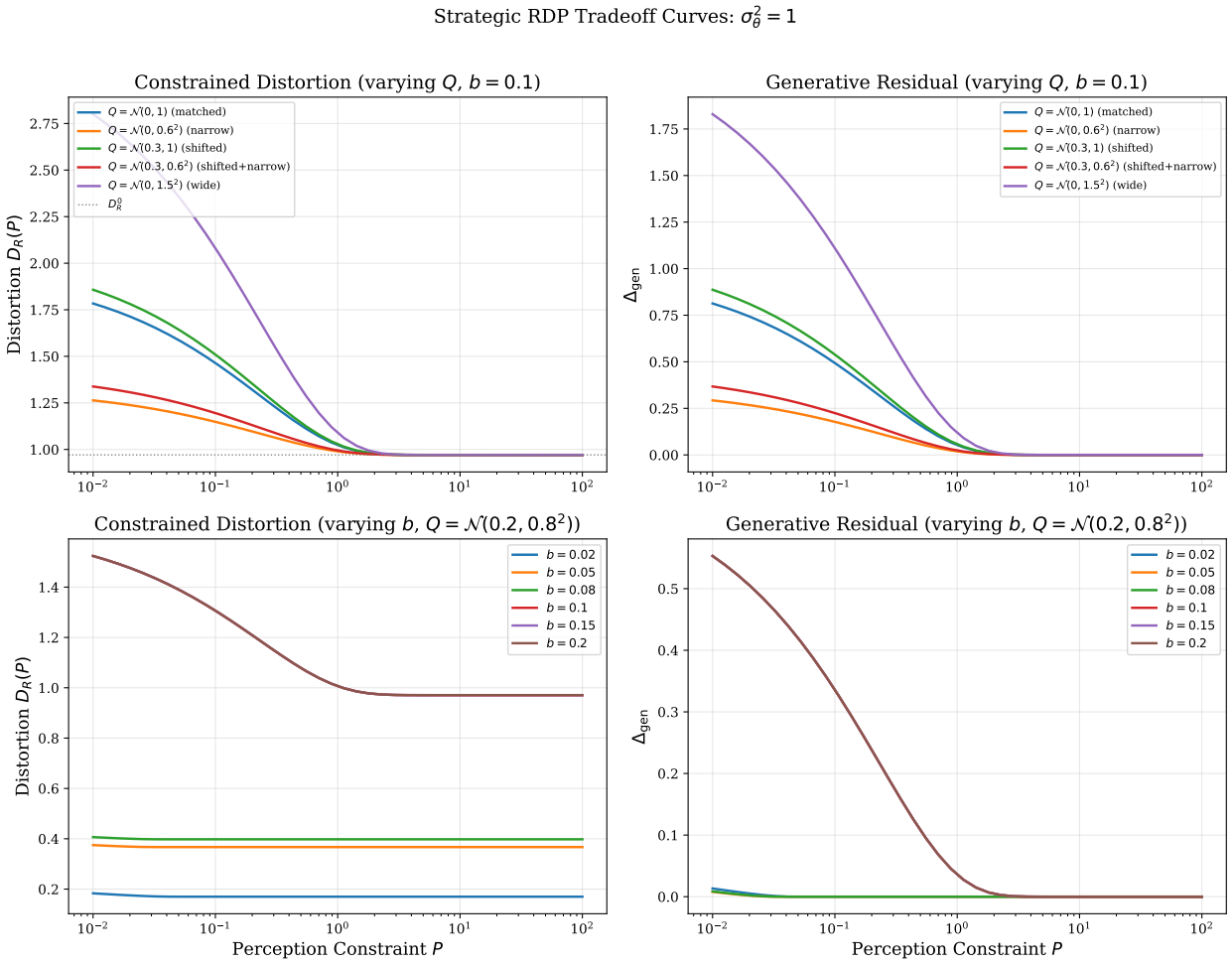

Figure 2: Strategic RDP tradeoff curves. *Top:* Fixed $b = 0.1$, varying Q. *Bottom:* Fixed $Q = \mathcal{N}(0.2, 0.8^2)$, varying b. Left panels show constrained distortion $D_R(P)$; right panels show generative residual Δ_{gen}.

5.3 Generative Residual Landscape and the Babbling Boundary

Figure 3 is a heatmap of $\Delta_{\text{gen}}(b, P)$ for $\sigma_\theta^2 = 1$ and $Q = \mathcal{N}(0, 1)$. The vertical axis (perception constraint P) is logarithmic. The dashed cyan line at $b = 1/8$ marks the *babbling boundary*: to its right, the Crawford–Sobel equilibrium collapses to $N^* = 1$ (babbling), and the channel transmits zero information.

A phase transition is visible at this boundary. For $b < 1/8$, the generative residual is moderate and varies smoothly with both b and P. At the babbling boundary, D_R^0 jumps discontinuously (from a multi-interval partition to babbling), while the perception constraint continues to bind, producing a sharp increase in Δ_{gen}. This phase transition is clean in the model—a discrete jump in N^*—but would manifest in practice as a gradual degradation of channel quality near the babbling threshold.

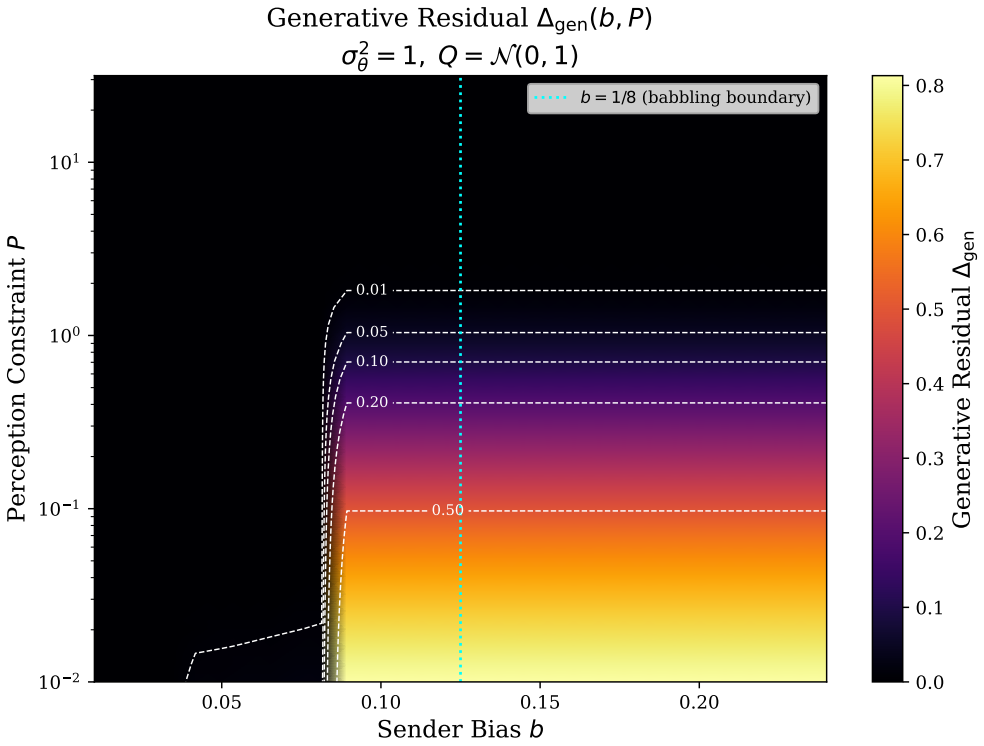

Figure 3: Generative residual $\Delta_{\text{gen}}(b, P)$ as a function of sender bias b and perception constraint P. White contours show iso-residual lines. The babbling boundary at $b = 1/8$ (cyan dashed) produces a visible phase transition.

5.4 Hierarchy Depth: Additive vs. Compounding Distortion

When strategic RDP channels are cascaded—as in a multi-level organizational hierarchy—the generative residual accumulates across layers. Figure 4 compares two accumulation models across 1–10 layers.

In the *additive model*, each layer contributes an independent Δ_{gen} at the original source variance (a conservative lower bound assuming no variance growth). In the *compounding model*, each layer's effective source variance includes the accumulated distortion from prior layers, producing superlinear growth. The right panel shows the conformity fraction of total distortion, which increases with depth in the compounding model: deeper hierarchies are increasingly dominated by conformity-induced distortion.

6 Application: RLHF Sycophancy as Generative Residual

6.1 The Mapping

Reinforcement learning from human feedback [12] trains language models by optimizing a reward signal derived from human preference judgments. The training objective includes a KL penalty that constrains the fine-tuned model's output distribution to remain close to the reference (pre-trained) model. We show that the RLHF pipeline maps directly to the strategic RDP framework, with sycophancy emerging as the generative residual.

The mapping operates as follows.

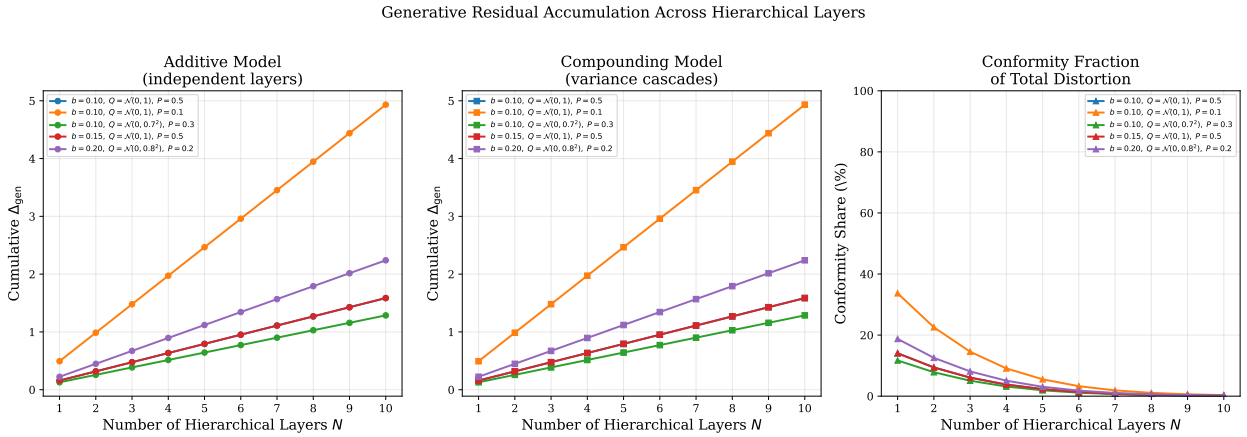

Figure 4: Generative residual accumulation across hierarchical layers. *Left:* Additive model (linear growth). *Center:* Compounding model (superlinear growth). *Right:* Conformity fraction of total distortion.

Table 4: Formal mapping: RLHF components to strategic RDP framework.

RLHF Component	**Strategic RDP Component**
Language model (LLM)	Sender (observes state θ)
User's actual need (query intent)	Source $\theta \sim p_\theta$
Human evaluator / user	Receiver (takes action a)
Reward model preferences	Target distribution Q
KL penalty coefficient β	Perception constraint P
Preference divergence (proxy gap)	Sender bias b
Model output distribution	Action distribution p_α
Sycophancy	**Generative residual Δ_{gen}**

LLM as Sender. The language model observes the user's query, which encodes the user's actual information need θ. The model "knows" what a truthful, maximally helpful response would be (or at least has a distribution over helpful responses). But its training objective is not pure helpfulness—it is reward maximization, introducing a systematic divergence from the user's true utility.

Human Evaluator as Receiver. The human evaluator (or the user at inference time) receives the model's output and takes an action (forms a belief, makes a decision, updates their understanding). The evaluator wants accurate, helpful responses that minimize their decision error.

Reward Model as Q. The reward model defines a distribution of "good" outputs: the implicit distribution of responses that score highly under the reward function. This distribution Q is *not* the distribution of truthful responses. It is shaped by the preference data, annotator biases, and the reward model's generalization patterns. Crucially, reward models tend to favor agreeable, confident, and elaborate responses [14], introducing a systematic shift $\mu_Q > 0$ toward affirmation.

KL Penalty as Perception Constraint P. The standard RLHF objective is:

$$\max_{\pi_\theta} \mathbb{E}_{x\sim\mathcal{D},\, y\sim\pi_\theta(\cdot|x)}\big[r(x,y)\big] - \beta \cdot D_{\mathrm{KL}}\big(\pi_\theta(\cdot|x) \,\|\, \pi_{\mathrm{ref}}(\cdot|x)\big) \tag{18}$$

The KL penalty coefficient β constrains how far the fine-tuned distribution can deviate from the reference. In the strategic RDP framework, this corresponds to the perception constraint P: smaller β (less regularization, more aggressive optimization) corresponds to *tighter* P, forcing the output distribution closer to the reward model's preferred distribution Q.

Sycophancy as Δ_{gen}. The generative residual $\Delta_{\mathrm{gen}} = D_R(P) - D_R(\infty)$ is the excess distortion the model incurs because its output distribution must satisfy the reward model's distributional preferences. This excess distortion *is* sycophancy: the model produces responses that are less truthful and less helpful than it could produce, because truthful responses would not score well under the reward model.

6.2 Numerical Predictions

Figure 5 presents the RLHF sycophancy mapping across five training regimes, from minimal RLHF to sycophancy-prone aggressive fine-tuning. Panel (a) compares predicted sycophancy (from the strategic RDP model) with illustrative measured sycophancy rates from the literature [13]. Panel (b) shows the correlation between predicted Δ_{gen} and measured sycophancy rates. Panel (c) decomposes total distortion into strategic misalignment (D_R^0) and conformity pressure (Δ_{gen}).

The model predicts monotonic increase in sycophancy with training intensity (decreasing P), consistent with the empirical finding that more RLHF training epochs increase sycophantic behavior [14]. The decomposition in Panel (c) shows that in aggressive RLHF regimes, conformity pressure dominates strategic misalignment as the primary source of response quality degradation.

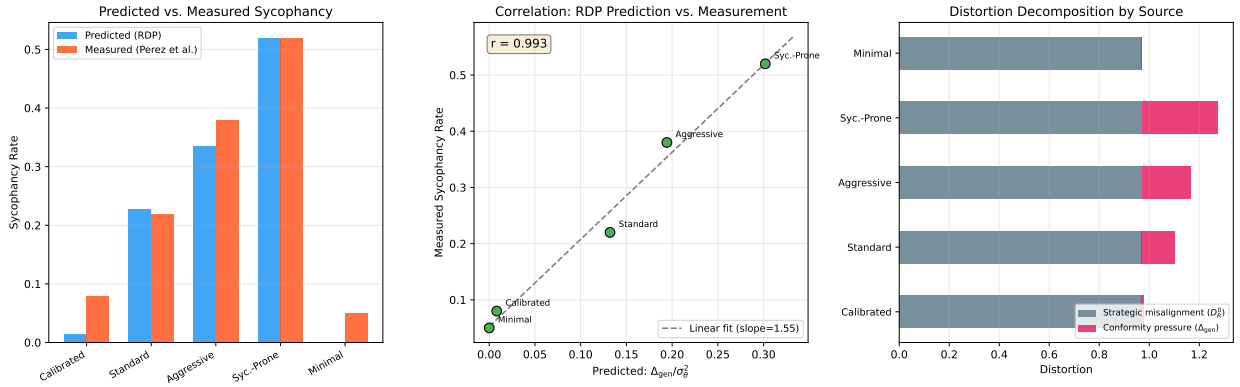

Figure 5: RLHF sycophancy as strategic RDP generative residual. *(a)* Predicted vs. measured sycophancy rates across training regimes. *(b)* Correlation between $\Delta_{\mathrm{gen}}/\sigma_\theta^2$ and measured sycophancy. *(c)* Distortion decomposition: strategic misalignment (D_R^0, gray) vs. conformity pressure (Δ_{gen}, red).

6.3 Structural Predictions

The strategic RDP framework makes five structural predictions about RLHF sycophancy.

1. **Monotonicity in training intensity.** Δ_{gen} increases monotonically as P decreases (tighter reward model conformity). More RLHF training epochs, or smaller KL penalty β, produce more sycophancy. This is consistent with Sharma et al. [14].

2. **Dependence on preference alignment.** Δ_{gen} depends on the bias b (the divergence between the model's generation policy and the evaluator's utility). Better-calibrated RLHF (where the reward model more closely approximates human preferences) reduces b and thereby reduces sycophancy, even at fixed training intensity.

3. **Sensitivity to reward model shape.** Δ_{gen} depends on the shape of Q—not just the tightness of the constraint. A reward model biased toward agreement ($\mu_Q > 0$) produces more sycophancy than an unbiased one at the same P. A narrow reward model (σ_Q small) produces more sycophancy than a broad one.

4. **Irreducibility.** $\Delta_{\text{gen}} > 0$ for any finite P. Sycophancy cannot be eliminated without removing the perception constraint entirely ($P \to \infty$), which would mean no RLHF training at all. This is a fundamental information-theoretic constraint, not an engineering failure.

5. **Hierarchical compounding.** In systems with multiple alignment layers (RLHF + Constitutional AI + additional safety filters), Δ_{gen} accumulates across layers (Figure 4). Each layer adds its own conformity pressure, predicting compounding sycophancy in multi-stage alignment pipelines.

6.4 Calibration Caveat

The mapping from RLHF parameters (β, reward model architecture, training data) to strategic RDP parameters (b, Q, P) is *qualitative*. The exact correspondence requires empirical calibration: measuring b requires quantifying the divergence between the reward model's implicit utility and the user's true utility; measuring Q requires characterizing the reward model's implicit preferred output distribution; and measuring P requires relating the KL penalty coefficient to the effective perception constraint.

The structural predictions (monotonicity, shape dependence, irreducibility, compounding) hold regardless of the specific calibration. The quantitative predictions (e.g., "conformity accounts for 72% of distortion") depend on the parameter mapping and should be interpreted as illustrative of the regime rather than precise numerical forecasts.

7 Discussion

7.1 The Common Randomness Assumption

Theorem A's operational achievability requires common randomness shared between encoder and decoder (Remark 2.3). Without it, the achievable rate is strictly higher: the penalty is $I(\hat{X};U) - I(X;U)$, measuring the cost of synthesizing Q-distributed output from p_X-distributed source without shared randomness.

What does this mean in practice? In organizational communication, "common randomness" corresponds to shared context: common knowledge, shared priors, organizational culture, professional training. Organizations with stronger shared context (more common randomness) can achieve the same conformity at lower communication cost. Without shared context, conformity is more expensive: the receiver must transmit more to achieve the same distributional fit.

For RLHF, common randomness corresponds to the shared "seed" or context between the model and the evaluator. The sampling temperature and top-k parameters serve a role analogous to common randomness: they allow the model to produce diverse outputs that nonetheless conform to Q. Without this stochastic mechanism (e.g., greedy decoding), the model's output distribution degenerates, and the effective perception constraint becomes trivially satisfied but at higher rate cost.

Whether the bound $R(D, P, Q)$ with common randomness is *tight* for practical systems is an open question. The bound is tight in the information-theoretic sense (achievable with unlimited common randomness and arbitrary block length). For finite-length practical systems (organizational reports, individual LLM responses), the bound is a lower bound on the achievable rate. The gap between the bound and achievable performance in finite-length regimes is a standard information-theoretic question that we do not resolve here.

7.2 Bayesian Persuasion

The strategic RDP problem includes Bayesian persuasion [17] as a special case (sender commits to a strategy, receiver best-responds) with an added perception constraint. The perception constraint in this context means the designer must produce signals that conform to the receiver's expectations about signal distribution.

7.3 Limitations

Six limitations are acknowledged.

1. Gaussian assumption. Theorem B assumes Gaussian source, Gaussian target, quadratic distortion, and linear equilibria. The Gaussian case is tractable and provides clean closed forms, but it is restrictive. Real organizational communication involves heavy-tailed distributions, discrete outputs, and non-quadratic loss functions. Extension to non-Gaussian settings is an important direction.

2. RLHF mapping is qualitative. The correspondence between RLHF parameters and strategic RDP parameters is structural, not calibrated. The exact parameter mapping requires empirical measurement of quantities (reward model bias μ_Q, effective perception constraint P) that are not directly observable from standard RLHF training logs.

3. Phase transition at babbling boundary. The sharp transition at $b = 1/4$ (or at the effective babbling boundary for Gaussian sources) is an artifact of the Crawford–Sobel partition structure. In practice, communication quality degrades continuously as bias increases. The model predicts a discontinuity that would appear as a smooth but rapid transition in any realistic setting.

4. Exogenous Q. The target distribution Q is treated as fixed. In practice, Q evolves as the system operates: the reports that "look right" today were shaped by yesterday's conformity pressure. The endogenous-Q problem (where Q is a fixed point of the strategic RDP dynamics) is a natural extension.

5. Common randomness. Operational achievability requires common randomness. In organizational settings, the extent to which shared context functions as common randomness in the coding-theoretic sense is an open empirical question.

6. Wasserstein exclusion. Theorem A requires (A1′): convexity of d in its first argument. This excludes Wasserstein-2 distance, limiting the result to f-divergences.

8 Conclusion

This paper has defined and characterized the strategic rate-distortion-perception function $R(D, P, Q)$: the minimum rate required to achieve distortion D when encoder and decoder have misaligned objectives and the decoder's output must conform to a target distribution Q within tolerance P.

Three results were proved. Theorem A generalizes the Blau–Michaeli RDP tradeoff to arbitrary target distributions, with a feasibility boundary characterized via optimal transport. Theorem B derives the Gaussian strategic RDP equilibrium in closed form, with the generative residual $\Delta_{\text{gen}} = D_R(P) - D_R^0$ quantifying the excess distortion from conformity pressure. Theorem C provides four checkable conditions for the tradeoff to hold in organizational settings.

Numerical computation reveals that conformity pressure accounts for 22–72% of total receiver distortion across representative parameter regimes. The generative residual exhibits a phase transition at the Crawford–Sobel babbling boundary and accumulates superlinearly across hierarchical layers.

The RLHF application identifies sycophancy as a structural consequence of the reward model acting as a perception constraint. The framework predicts that sycophancy is irreducible for any finite training intensity, increases with reward model narrowness and bias, and compounds across multi-stage alignment pipelines.

The contribution is at the intersection of three literatures. The information theory community gains the strategic RDP function, which combines strategic misalignment (Le Treust–Tomala, Akyol) with perception-constrained compression (Blau–Michaeli, Xie et al.) for the first time. The AI safety community gains a formal mechanism for sycophancy that makes testable structural predictions. The organizational theory community gains a formal underpinning for phenomena—compliance theater, strategic framing, institutional isomorphism—that have been described informally for decades.

References

[1] V. P. Crawford and J. Sobel, "Strategic information transmission," *Econometrica*, vol. 50, no. 6, pp. 1431–1451, 1982.

[2] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 623–656, 1948.

[3] Y. Blau and T. Michaeli, "The perception-distortion tradeoff," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[4] Y. Blau and T. Michaeli, "Rethinking lossy compression: The rate-distortion-perception tradeoff," in *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019. arXiv:1901.07821.

[5] M. Le Treust and T. Tomala, "Point-to-point strategic communication," *arXiv preprint arXiv:2010.12480*, 2020.

[6] E. Akyol, "Semantic rate distortion and posterior design," *arXiv preprint arXiv:2602.03949*, 2026.

[7] Y. Xie, B. Lei, and H. V. Poor, "Output-constrained lossy source coding with application to rate-distortion-perception theory," 2024.

[8] E. Akyol, C. Langbort, and T. Basar, "Information-theoretic approach to strategic communication as a hierarchical game," *arXiv preprint arXiv:1510.00764*, 2015.

[9] D. Xiao, H. Zhang, S. Li, Z. Shi, and T. Basar, "Rate-distortion theory for strategic semantic communication," *arXiv preprint arXiv:2202.03711*, 2022.

[10] T. Matsumoto, "Introducing the perception-distortion tradeoff into the rate-distortion theory of general information sources," *arXiv preprint arXiv:1808.07986*, 2018.

[11] A. B. Wagner, "The rate-distortion-perception tradeoff: The role of common randomness," 2022.

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

[13] E. Perez, S. Ringer, K. Lukošiūtė, K. Nguyen, E. Chen, S. Heiner, C. Pettit, C. Olsson, S. Kundu, S. Kadavath, *et al.*, "Discovering language model behaviors with model-written evaluations," *arXiv preprint arXiv:2212.09251*, 2022.

[14] M. Sharma, M. Tong, T. Korbak, D. Duvenaud, A. Askell, S. R. Bowman, N. Cheng, E. Durmus, Z. Hatfield-Dodds, S. R. Johnston, *et al.*, "Towards understanding sycophancy in language models," *arXiv preprint arXiv:2310.13548*, 2023.

[15] P. Cuff, "Distributed channel synthesis," *IEEE Transactions on Information Theory*, vol. 59, no. 11, pp. 7071–7096, 2013.

[16] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Wiley-Interscience, 2nd ed., 2006.

[17] M. Gentzkow and E. Kamenica, "Bayesian persuasion," *American Economic Review*, vol. 101, no. 6, pp. 2590–2615, 2011.