Knowledge Divergence and the Value of Debate for Scalable Oversight

Anonymous ACL submission

Abstract

AI safety via debate and reinforcement learning from AI feedback (RLAIF) are both proposed methods for scalable oversight of language models, yet no formal framework relates them or characterizes when debate offers an advantage. We analyze this by parameterizing debate's value through the geometry of knowledge divergence between debating models. Using principal angles between models' representation subspaces, we prove that the debate advantage admits an exact closed form. When models share identical training corpora, debate reduces to RLAIF-like where a single-agent method recovers the same optimum. When models possess divergent knowledge, debate advantage scales with a phase transition from quadratic regime (debate offers negligible benefit) to linear regime (debate is essential). We classify three regimes of knowledge divergence (shared, one-sided, and compositional) and provide existence results showing that debate can achieve outcomes inaccessible to either model alone, alongside a negative result showing that sufficiently strong adversarial incentives cause coordination failure in the compositional regime, with a sharp threshold separating effective from ineffective debate. We propose a formal connection between debate and RLAIF, a geometric foundation for understanding when adversarial oversight protocols are justified, and connection to the problem of eliciting latent knowledge across models with complementary information.

1 Introduction

Scalable oversight, namely the problem of supervising AI systems on tasks too complex for direct human evaluation, has motivated two prominent families of approaches. AI safety via debate (Irving et al., 2018; Brown-Cohen et al., 2024) proposes pitting two AI models against each other in a structured argument, with a human judge evaluating the transcript. Reinforcement learning from AI feedback (RLAIF), as implemented in Constitutional AI (Bai et al., 2022), trains models to self-critique against a set of principles. Both approaches aim to amplify a weak overseer's ability to evaluate complex model outputs.

Despite their shared motivation, these methods have developed in isolation. The debate literature, rooted in computational complexity theory, studies the power of interactive proof systems with polynomial-time verifiers (Brown-Cohen et al., 2024, 2025). The RLAIF literature, rooted in preference learning, studies how constitutional principles shape model distributions. No formal framework relates these approaches or characterizes when one offers an advantage over the other.

We analyze this question. Under our assumptions, our central observation is that the value of debate depends on the knowledge divergence between the debating models, which is a quantity that existing debate theory, which treats provers as abstract computational agents, does not formalize. Using principal angles between models' representation subspaces, we derive a bound on the debate advantage as the improvement in constitutional score achievable through adversarial multi-model interaction over single-model optimization.

Our contributions are as follows. First, we prove that the debate advantage admits the exact closed form $\Delta = \sqrt{(K_A^*)^2 + \eta^2} - K_A^*$ and satisfies tight bounds governed by a private information value η derived from the principal angle spectrum between models' representation subspaces (Theorem 6). Same-corpus equivalence falls out; when models share training data, $\eta = 0$ and debate reduces to RLAIF (Corollary 7). Then, we classify three regimes of knowledge divergence (shared, one-sided, and compositional) with existence results showing debate can achieve outcomes inaccessible to either model alone (Propositions 15 and 16), and a negative result showing that ad-

versarial incentives above a sharp threshold cause coordination failure in the compositional regime (Proposition 17).

We provide a theoretical explanation for the empirical finding that model homogeneity undermines AI oversight (Goel et al., 2025), predicts that debate's value scales with knowledge diversity, and suggests that the interesting regime for debate which concerns knowledge-divergent models is the regime that is currently understudied.

2 Geometric Framework

2.1 Setup and Notation

Let $\mathcal{Y}$ be a finite output space and let $h : \mathcal{Y} \to \mathbb{R}^d$ be a representation map embedding outputs into a d-dimensional space. Two models A and B, trained on (potentially different) corpora, induce k-dimensional representation subspaces $V_A, V_B \subseteq \mathbb{R}^d$. We assume $\dim(V_A) = \dim(V_B) = k$ for simplicity; the framework extends to unequal dimensions.

Definition 1 (Principal Angles)**.** *The principal angles $\theta_1, \ldots, \theta_k \in [0, \pi/2]$ between V_A and V_B are defined recursively by*

$$\cos\theta_i = \max_{\substack{u \in V_A, \|u\|=1 \\ u \perp u_1, \ldots, u_{i-1}}} \max_{\substack{v \in V_B, \|v\|=1 \\ v \perp v_1, \ldots, v_{i-1}}} \langle u, v \rangle, \quad (1)$$

where (u_i, v_i) are the corresponding principal vectors *satisfying $\langle u_i, v_j \rangle = \cos\theta_i \cdot \delta_{ij}$.*

The principal angles provide a characterization of the relative geometry of the two subspaces. When $\theta_i = 0$ for all i, the subspaces are identical; when $\theta_i = \pi/2$ for all i, they are orthogonal.

We model the constitutional scoring function as a linear functional in representation space.

Definition 2 (Linear Constitutional Scoring)**.** *A constitutional scoring function is $K(y) = \langle w, h(y) \rangle$ for a fixed preference direction $w \in \mathbb{R}^d$ with $\|w\| = 1$.*

The linear assumption is standard in the representation geometry literature (Park et al., 2024; Arditi et al., 2024; Turner et al., 2024; Rimsky et al., 2024) and permits a clean geometric analysis. We discuss relaxation to nonlinear K in Section 5.

2.2 RLAIF and Debate as Optimization

Under linear scoring, each model's RLAIF-optimal constitutional score is determined by how much of the preference direction w its representation subspace captures:

$$K_A^* = \|\Pi_{V_A} w\|, \qquad K_B^* = \|\Pi_{V_B} w\|, \quad (2)$$

where Π_V denotes orthogonal projection onto V.

Debate between models A and B pools their representational knowledge. At equilibrium, debate can access any output representable in either model's subspace, so the debate-optimal score is:

$$K_{AB}^* = \|\Pi_{V_A + V_B} w\|, \quad (3)$$

where $V_A + V_B = \{u + v : u \in V_A, v \in V_B\}$ is the Minkowski sum.

Definition 3 (Debate Advantage)**.** *The* debate advantage *is*

$$\Delta(V_A, V_B, w) = K_{AB}^* - \max(K_A^*, K_B^*). \quad (4)$$

The debate advantage measures the improvement in constitutional score achievable by pooling two models' knowledge through adversarial interaction, over the best that either model achieves alone.

2.3 Decomposition via Principal Vectors

We decompose the combined subspace $V_A + V_B$ to isolate the contribution of private knowledge. For each principal angle $\theta_i > 0$, define the private direction:

$$\tilde{v}_i = \frac{v_i - \cos\theta_i \cdot u_i}{\sin\theta_i}. \quad (5)$$

These are the components of V_B's principal vectors orthogonal to V_A, normalized.

Lemma 4. *The set $\{u_1, \ldots, u_k, \tilde{v}_1, \ldots, \tilde{v}_m\}$, where $m = |\{i : \theta_i > 0\}|$, forms an orthonormal basis for $V_A + V_B$.*

Proof. That each $\tilde{v}_i$ has unit norm follows from $\|v_i - \cos\theta_i \cdot u_i\|^2 = 1 - \cos^2\theta_i = \sin^2\theta_i$. For orthogonality, $\langle \tilde{v}_i, u_j \rangle = (\langle v_i, u_j \rangle - \cos\theta_i \cdot \delta_{ij}) / \sin\theta_i = 0$ by the principal angle property. For $i \neq j$:

$$\langle \tilde{v}_i, \tilde{v}_j \rangle = \frac{1}{\sin\theta_i \sin\theta_j} \big(\langle v_i, v_j \rangle - \cos\theta_j \langle v_i, u_j \rangle - \cos\theta_i \langle u_i, v_j \rangle + \cos\theta_i \cos\theta_j \langle u_i, u_j \rangle \big) \quad (6)$$

$$= \frac{\delta_{ij} - 0 - 0 + 0}{\sin\theta_i \sin\theta_j} = \delta_{ij}. \quad (7)$$

Spanning follows because each $v_i = \cos\theta_i \cdot u_i + \sin\theta_i \cdot \tilde{v}_i \in \mathrm{span}\{u_i, \tilde{v}_i\}$. □

Using this basis, we decompose the projection of w onto $V_A + V_B$:

$$\|\Pi_{V_A+V_B} w\|^2 = \underbrace{\sum_{i=1}^{k} \langle w, u_i \rangle^2}_{=\|\Pi_{V_A} w\|^2} + \underbrace{\sum_{i:\theta_i>0} \langle w, \tilde{v}_i \rangle^2}_{=:\eta^2}. \tag{8}$$

Definition 5 (Private Information Value)**.** *The private information value of model B relative to model A under preference direction w is*

$$\eta(V_A, V_B, w) = \sqrt{\sum_{i:\theta_i>0} \langle w, \tilde{v}_i \rangle^2}, \tag{9}$$

where $\langle w, \tilde{v}_i \rangle = (\langle w, v_i \rangle - \cos\theta_i \cdot \langle w, u_i \rangle)/\sin\theta_i$.

The private information value measures K-relevant information residing in directions V_B contributes beyond V_A. The numerator captures component of w along v_i that is not already explained by v_i's projection onto u_i and the denominator normalizes by the geometric novelty of each direction.

2.4 Main Result

Theorem 6 (Debate Advantage Bound)**.** *Under linear constitutional scoring $K(y) = \langle w, h(y) \rangle$, and assuming without loss of generality that $K_A^* \geq K_B^*$, debate advantage admits the exact closed form*

$$\Delta = \sqrt{(K_A^*)^2 + \eta^2} - K_A^*, \tag{10}$$

and satisfies the bounds

$$\frac{\eta^2}{2K_A^* + \eta} \leq \Delta \leq \eta. \tag{11}$$

Both bounds are tight. the lower bound is achieved as $\eta \to 0$; the upper bound is achieved when $K_A^* = 0$.

Proof. By Lemma 4 and the decomposition of $\Pi_{V_A+V_B} w$:

$$K_{AB}^* = \|\Pi_{V_A+V_B} w\| = \sqrt{(K_A^*)^2 + \eta^2}. \tag{12}$$

The closed form (10) follows immediately from Definition 3.

Upper bound. For $a, b \geq 0$, $\sqrt{a^2 + b^2} \leq a + b$, so $\Delta = \sqrt{(K_A^*)^2 + \eta^2} - K_A^* \leq \eta$. Equality holds when $K_A^* = 0$, i.e. $w \perp V_A$.

Lower bound. Rationalizing the exact expression:

$$\Delta = \frac{(K_A^*)^2 + \eta^2 - (K_A^*)^2}{\sqrt{(K_A^*)^2 + \eta^2} + K_A^*} = \frac{\eta^2}{K_{AB}^* + K_A^*}. \tag{13}$$

Since $K_{AB}^* = \sqrt{(K_A^*)^2 + \eta^2} \leq K_A^* + \eta$ (as $(K_A^*)^2 + \eta^2 \leq (K_A^* + \eta)^2$ for $K_A^*, \eta \geq 0$), we obtain

$$\Delta \geq \frac{\eta^2}{(K_A^* + \eta) + K_A^*} = \frac{\eta^2}{2K_A^* + \eta}. \tag{14}$$

For tightness: as $\eta \to 0$, $\Delta = \eta^2/(2K_A^*) + O(\eta^4)$ while $\eta^2/(2K_A^* + \eta) = \eta^2/(2K_A^*) + O(\eta^3)$, so the bound is asymptotically tight. □

Remark 1 (Scaling regimes)**.** *The rationalized form $\Delta = \eta^2/(K_{AB}^* + K_A^*)$ reveals two qualitative regimes:*

- ***Small private information*** *($\eta \ll K_A^*$): $K_{AB}^* \approx K_A^*$, so $\Delta \approx \eta^2/2K_A^*$. The debate advantage is quadratically small in the private information value. The overhead of running a multi-model debate protocol is unlikely to be justified.*
- ***Large private information*** *($\eta \gg K_A^*$): $K_{AB}^* \approx \eta$, so $\Delta \approx \eta$. The debate advantage is linear in the private information value. Debate is essential as single-model optimization misses most of the achievable constitutional score.*

The transition between regimes occurs at $\eta \approx K_A^*$, i.e. when the K-relevant private information is comparable to the K-relevant shared information.

2.5 Corollaries

Corollary 7 (Same-Corpus Equivalence)**.** *If $V_A = V_B$ (i.e. models share representation subspaces), then $\theta_i = 0$ for all i, $\eta = 0$, and $\Delta = 0$. Debate and RLAIF achieve identical constitutional scores.*

Proof. When $\theta_i = 0$ for all i, there are no private directions: $V_A + V_B = V_A$. Thus $\eta = 0$ and $K_{AB}^* = K_A^*$. □

Remark 2. *Under our formalization RLAIF can be viewed as depth-1 debate under the same-corpus assumption. The model evaluates its own output against constitutional principles, which is a single round of self-debate. Corollary 7 shows this is without loss of optimality when knowledge is shared.*

Corollary 8 (Maximum Divergence)**.** *If $V_A \perp V_B$ (i.e. $\theta_i = \pi/2$ for all i), then $\tilde{v}_i = v_i$, $\eta = \|\Pi_{V_B} w\| = K_B^*$, and $\Delta = \sqrt{(K_A^*)^2 + (K_B^*)^2} - K_A^*$.*

Corollary 9 (Isotropic Scaling)**.** *If w has equal projection onto each principal direction pair, i.e., $\langle w, u_i\rangle = \langle w, v_i\rangle = \alpha$ for all i and all principal angles equal θ, then*

$$\eta = \alpha\sqrt{k} \cdot |\tan(\theta/2)|, \tag{15}$$

and the debate advantage transitions from $\Delta \sim k\alpha^2 \tan^2(\theta/2)/(2K_A^*)$ for small θ to $\Delta \sim \alpha\sqrt{k} \cdot \tan(\theta/2)$ for large θ.

Proof. Under the isotropic assumption, $\langle w, \tilde{v}_i\rangle = \alpha(1-\cos\theta)/\sin\theta = \alpha\tan(\theta/2)$ for each i. Thus $\eta = \alpha\sqrt{k} \cdot \tan(\theta/2)$. The scaling regimes follow from the exact form (10) and Remark 1. □

This corollary provides a summary of the debate advantage, which scales as $\tan(\theta/2)$ in the principal angles between models' representation subspaces. The tangent half-angle captures a phase transition where for small angles (similar models), debate advantage is negligible. As angles approach $\pi/2$ (complementary models), debate advantage grows without bound.

2.6 Multi-Agent Debate

The two-model framework extends naturally to n debaters. Let $A_1, \ldots, A_n$ be models with representation subspaces $V_{A_1}, \ldots, V_{A_n} \subseteq \mathbb{R}^d$. Define the cumulative subspace

$$S_j = V_{A_1} + \cdots + V_{A_j}, \qquad S_0 = \{0\},$$

the n-model debate-optimal score $K^*_{A_1\cdots A_n} = \|\Pi_{S_n} w\|$, and the n-model debate advantage $\Delta_n = K^*_{A_1\cdots A_n} - \max_i K^*_{A_i}$.

Definition 10 (Marginal Private Information Value)**.** *The marginal private information value of model A_j given the coalition $\{A_1, \ldots, A_{j-1}\}$ is*

$$\eta_j^2 = \|\Pi_{S_j} w\|^2 - \|\Pi_{S_{j-1}} w\|^2.$$

Proposition 11 (Multi-Agent Decomposition)**.** *The n-model debate-optimal score satisfies*

$$K^*_{A_1\cdots A_n} = \sqrt{(K^*_{A_{\sigma(1)}})^2 + \sum_{j=2}^{n} \eta^2_{\sigma(j)}} \tag{16}$$

for any ordering σ of the models, where $\eta_{\sigma(j)}$ is the marginal private information value of $A_{\sigma(j)}$ relative to S_{j-1}.

Proof. Telescoping:

$$\|\Pi_{S_n} w\|^2 = \|\Pi_{S_1} w\|^2 + \sum_{j=2}^{n} \big(\|\Pi_{S_j} w\|^2 - \|\Pi_{S_{j-1}} w\|^2\big) \tag{17}$$

$$= (K^*_{A_{\sigma(1)}})^2 + \sum_{j=2}^{n} \eta^2_{\sigma(j)}. \tag{18}$$

This is valid because $\Pi_{S_j} - \Pi_{S_{j-1}}$ is the orthogonal projection onto $S_j \cap S_{j-1}^{\perp}$, and these projections are mutually orthogonal for different j. □

Proposition 12 (Diminishing Marginal Returns)**.** *The marginal debate advantage of adding model A_j to coalition $\{A_1, \ldots, A_{j-1}\}$,*

$$\delta_j = \sqrt{\|\Pi_{S_{j-1}} w\|^2 + \eta_j^2} - \|\Pi_{S_{j-1}} w\|, \tag{19}$$

has the same functional form as the two-model debate advantage (Theorem 2.6) and satisfies the same bounds. Under the greedy ordering (selecting at each step the model maximizing η_j):

(i) The marginal debate advantage δ_j is non-increasing in j.

(ii) For constant marginal private information $\eta_j = \eta$, the function $f(a) = \sqrt{a^2+\eta^2} - a$ is strictly decreasing in a, and since $\|\Pi_{S_{j-1}} w\|$ is non-decreasing, δ_j is strictly decreasing.

(iii) The total private information is bounded: $\sum_{j=1}^{n} \eta_j^2 \le 1$.

Proof. The marginal advantage has the form of Theorem 2.6 with K_A^* replaced by $\|\Pi_{S_{j-1}} w\|$. Part (i): $\|\Pi_{S_{j-1}} w\|$ is non-decreasing in j (subspaces grow), and $f(a) = \sqrt{a^2+\eta^2} - a$ has $f'(a) = a/\sqrt{a^2+\eta^2} - 1 < 0$. Part (ii) follows from strict monotonicity of f when $\eta > 0$. Part (iii): $\sum_j \eta_j^2 = \|\Pi_{S_n} w\|^2 - (K^*_{A_{\sigma(1)}})^2 \le \|w\|^2 = 1$. □

Corollary 13 (Optimal Coalition Size)**.** *If each additional debater incurs cost $c > 0$, the net value of the j-th debater is $\delta_j - c$. In the small-η regime ($\delta_j \approx \eta^2/(2\|\Pi_{S_{j-1}} w\|)$), debate should stop adding models once*

$$\|\Pi_{S_{j-1}} w\| \ge \frac{\eta^2}{2c},$$

i.e. once the cumulative captured projection is large enough that the next model's marginal contribution falls below cost.

Remark 3. *The self-similarity of the marginal advantage where each new model contributes via the same ($\sqrt{a^2+b^2}-a$) formula means the two-model theory is the complete building block for the multi-agent case. The new content is the diminishing returns structure, which provides a concrete stopping rule for multi-agent debate protocols.*

3 Discrete Regimes of Knowledge Divergence

The geometric bound of Theorem 6 admits a natural interpretation through three discrete regimes, determined by where the preference direction w loads relative to the principal subspaces.

Proposition 14 (Regime Classification)**.** *Let $w_A = \Pi_{V_A} w$, $w_B = \Pi_{V_B} w$, and $w_\perp = w - \Pi_{V_A+V_B} w$. The debate advantage satisfies:*

- ***(i) Shared knowledge.*** *If $\Pi_{V_A} w = \Pi_{V_A+V_B} w$ (i.e. $\eta = 0$), then $\Delta = 0$. The K-optimal output is accessible to both models.*
- ***(ii) One-sided private knowledge.*** *If $\eta > 0$ and w loads on private directions from only one model (say $\langle w, \tilde{v}_i \rangle > 0$ for some i), then $\Delta > 0$. The better output is known to one debater but not the other.*
- ***(iii) Compositional private knowledge.*** *If achieving K^*_{AB} requires combining representations from $V_A \setminus V_B$ and $V_B \setminus V_A$, then $\Delta > 0$ but realizability depends on the debate protocol.*

3.1 One-Sided Revelation

Proposition 15 (One-Sided Revelation)**.** *There exist representation subspaces V_A, V_B and preference direction w such that:*

1. *$K^*_{AB} > K^*_A \geq K^*_B$;*
2. *At Nash equilibrium of the debate game, model A reveals the K-optimal output;*
3. *Neither RLAIF applied to A nor RLAIF applied to B alone achieves K^*_{AB}.*

Proof (Construction). Let $d = 3$, $k = 2$. Let $V_A = \text{span}\{e_1, e_2\}$ and $V_B = \text{span}\{e_1, e_3\}$. The principal angles are $\theta_1 = 0$ (shared direction e_1) and $\theta_2 = \pi/2$ (private directions e_2, e_3).

Let $w = \frac{1}{\sqrt{2}}(e_2 + e_3)$. Then $K^*_A = \|\Pi_{V_A} w\| = 1/\sqrt{2}$ and $K^*_B = \|\Pi_{V_B} w\| = 1/\sqrt{2}$, while $K^*_{AB} = \|w\| = 1$. Thus $\Delta = 1 - 1/\sqrt{2} > 0$.

In the debate game, model A can argue for the output with representation e_2 and model B for e_3. The judge, observing both, can score the combination $\frac{1}{\sqrt{2}}(e_2 + e_3)$ higher than either alone. Model A's private direction contributes K-relevant information that B cannot access, and vice versa; the debate structure forces both to reveal their private knowledge because withholding it means losing to the opponent's argument for their own private-knowledge output. □

3.2 Compositional Existence

Proposition 16 (Compositional Existence)**.** *There exist V_A, V_B, w, and a debate protocol such that the debate outcome achieves a constitutional score that requires features from both $V_A \setminus V_B$ and $V_B \setminus V_A$, and that neither model can achieve alone.*

Proof (Construction). Let $d = 4$, $k = 2$. Let $V_A = \text{span}\{e_1, e_2\}$ and $V_B = \text{span}\{e_3, e_4\}$, so $V_A \perp V_B$ and all principal angles are $\pi/2$.

Define a nonlinear[1] constitutional scoring function that rewards outputs combining features from both subspaces: $K(y) = \min(\langle h(y), e_1 \rangle, \langle h(y), e_3 \rangle)$. This function is maximized only by outputs with positive projection onto both $e_1 \in V_A$ and $e_3 \in V_B$.

Model A's RLAIF optimizes K over $\text{supp}(P_A) \subseteq V_A$: since $\langle h(y), e_3 \rangle = 0$ for all y representable by A, we have $K^*_A = 0$. Similarly, $K^*_B = 0$.

In a two-round debate protocol:

Round 1: Model A proposes an output with maximal $\langle h(y), e_1 \rangle$.

Round 2: Model B, observing A's proposal, augments it with e_3 to produce a combined output. The judge scores $K = \min(\langle h, e_1 \rangle, \langle h, e_3 \rangle) > 0 = \max(K^*_A, K^*_B)$. □

3.3 Limits of Adversarial Debate

The previous proposition shows that debate protocols can achieve compositional advantage. However, debate protocols are adversarial by design as each debater has an incentive to "win" the debate, not merely to produce the best output. We show that sufficiently strong adversarial incentives, which are the same incentives that make debate use-

[1] We briefly relax linearity for this construction. The existence claim extends to a piecewise-linear approximation compatible with the overall linear model.

ful for oversight, can prevent debate from realizing the compositional optimum.

Proposition 17 (Adversarial Coordination Failure)**.** *There exist knowledge-divergent debate games with $\Delta > 0$ and adversarial incentive parameter λ such that:*

1. *For $\lambda < \lambda^*$, the subgame-perfect equilibrium (SPE) achieves K^*_{AB}.*
2. *For $\lambda > \lambda^*$, the unique SPE achieves constitutional score strictly below K^*_{AB}.*

The threshold is $\lambda^* = K^*_{AB} - K^*_{\text{safe}}$, the gap between the compositional optimum and the best score achievable without inter-model coordination.

Proof. Let $V_A = \text{span}\{e_1\}$ and $V_B = \text{span}\{e_2\}$ with $V_A \perp V_B$. Consider a two-action debate where each model chooses either a generous action g (contributing its best private information for composition) or a strategic action s (optimizing its individual position). Constitutional scores are:

$$K(g,g) = R, K(s,s) = P, K(g,s) = K(s,g) = S, \tag{20}$$

with $R > P > S \geq 0$. Here $R = K^*_{AB}$ is the compositional optimum requiring both models' contributions, $P = K^*_{\text{safe}}$ is the score from mutual conservative play, and S is the mismatch penalty when one model contributes generously while the other plays strategically.

Each debater's payoff augments the constitutional score with an adversarial bonus $\lambda \geq 0$ for playing s:

$$U_A(a_A, a_B) = K(a_A, a_B) + \lambda \cdot \mathbb{1}[a_A = s], \tag{21}$$

$$U_B(a_A, a_B) = K(a_A, a_B) + \lambda \cdot \mathbb{1}[a_B = s]. \tag{22}$$

The adversarial bonus models the competitive incentive in debate: playing strategically (arguing forcefully for one's own position) provides a payoff advantage independent of the joint output quality.

Backward induction. In the sequential game, model A moves first and model B observes a_A before choosing a_B.

If $a_A = g$: Model B compares $U_B(g,g) = R$ with $U_B(g,s) = S + \lambda$. B plays g if and only if $R > S + \lambda$, i.e., $\lambda < R - S$.

If $a_A = s$: Model B compares $U_B(s,g) = S$ with $U_B(s,s) = P + \lambda$. Since $P > S$ and $\lambda \geq 0$, B always plays s.

Model A anticipates B's strategy. When $\lambda < R - S$ (so B responds g to g), A compares $U_A(g,g) = R$ with $U_A(s,s) = P + \lambda$, and plays g iff $\lambda < R - P$. Since $R - P < R - S$ (as $P > S$), three regimes emerge:

(a) $\lambda < R - P = \lambda^*$: The SPE is (g,g) with $K = R = K^*_{AB}$.

(b) $R - P < \lambda < R - S$: B would cooperate, but A defects. The SPE is (s,s) with $K = P < K^*_{AB}$.

(c) $\lambda > R - S$: Both models defect regardless. The SPE is (s,s) with $K = P < K^*_{AB}$.

For $\lambda > \lambda^* = R - P$, the unique SPE achieves $K = P = K^*_{\text{safe}} < R = K^*_{AB}$. The gap is $R - P = \Delta$, the full debate advantage. □

Remark 4. *Proposition 17 reveals a phase transition in debate effectiveness. Below the threshold $\lambda^* = K^*_{AB} - K^*_{\text{safe}}$, adversarial incentives are compatible with composition as the coordination benefit exceeds the temptation to defect, and the SPE achieves the full debate advantage. Above the threshold, the adversarial incentive overwhelms coordination, and debate collapses to the non-compositional outcome.*

Regime (b) is particularly interesting. Model B would cooperate if A led with the generous action, but A defects because the temptation payoff $P + \lambda$ exceeds the coordination payoff R. The sequential structure of debate which is essential for the honest revelation of Proposition 15 becomes a liability in the compositional setting by giving the first mover an incentive to deviate.

The threshold λ^* characterizes when debate's adversarial structure is "too strong" for composition when the competitive incentive exceeds the gap between compositional and safe outcomes. This suggests that compositional knowledge aggregation may require weaker adversarial incentives than those designed for oversight or cooperative structures altogether (Chen et al., 2025).

3.4 Dynamic Subspaces Under Debate

The preceding analysis treats representation subspaces as fixed. In practice, models may update their effective representations during debate through in-context learning. We formalize this as a T-round process in which subspaces evolve.

At round t, models have subspaces $V_A^{(t)}$ and $V_B^{(t)}$, initialized at $V_A^{(0)} = V_A$ and $V_B^{(0)} = V_B$. The

private information value $\eta^{(t)}$ and debate advantage $\Delta^{(t)}$ are defined as in Section 2 with these time-indexed subspaces. A key simplification: absorbing a direction from $V_B^{(t)}$ into $V_A^{(t)}$ does not expand the combined subspace $V_A^{(t)} + V_B^{(t)}$, since the absorbed direction already belonged to it. Therefore K_{AB}^* is constant across rounds, and the debate advantage simplifies to

$$\Delta^{(t)} = K_{AB}^* - K_A^{*(t)}. \tag{23}$$

Cooperative dynamics.

Under cooperative revelation, model B reveals at each round the private direction most aligned with w:

$$d_B^{(t)} = \arg \max_{\substack{d \in V_B^{(t)},\, \|d\|=1 \\ d \perp V_A^{(t)}}} |\langle w, d \rangle|,$$

and model A absorbs it: $V_A^{(t+1)} = V_A^{(t)} \oplus \mathrm{span}\{d_B^{(t)}\}$.

Proposition 18 (Convergence Under Cooperative Dynamics)**.** *Under cooperative revelation:*

(i) $K_A^{*(t)}$ *is strictly increasing whenever* $\eta^{(t)} > 0$.

(ii) The private information value decreases as

$$(\eta^{(t+1)})^2 = (\eta^{(t)})^2 - \langle w, d_B^{(t)} \rangle^2. \tag{24}$$

(iii) The per-round decrease in debate advantage is

$$\Delta^{(t)} - \Delta^{(t+1)} = K_A^{*(t+1)} - K_A^{*(t)} = \frac{\langle w, d_B^{(t)} \rangle^2}{K_A^{*(t+1)} + K_A^{*(t)}}. \tag{25}$$

(iv) Full convergence occurs in at most m *rounds, where* $m = |\{i : \theta_i^{(0)} > 0\}|$*:*

$$\eta^{(T)} = 0 \ \textit{for all} \ T \geq m, \qquad K_A^{*(m)} = K_{AB}^*.$$

Proof. (i) Absorbing $d_B^{(t)}$ gives $(K_A^{*(t+1)})^2 = (K_A^{*(t)})^2 + \langle w, d_B^{(t)} \rangle^2$. The greedy selection ensures $\langle w, d_B^{(t)} \rangle^2 > 0$ whenever $\eta^{(t)} > 0$.

(ii) Since K_{AB}^* is constant: $(\eta^{(t)})^2 = (K_{AB}^*)^2 - (K_A^{*(t)})^2$, so $(\eta^{(t+1)})^2 = (K_{AB}^*)^2 - (K_A^{*(t)})^2 - \langle w, d_B^{(t)} \rangle^2 = (\eta^{(t)})^2 - \langle w, d_B^{(t)} \rangle^2$.

(iii) $\Delta^{(t)} - \Delta^{(t+1)} = K_A^{*(t+1)} - K_A^{*(t)} = \sqrt{(K_A^{*(t)})^2 + \langle w, d_B^{(t)} \rangle^2} - K_A^{*(t)}$. Rationalizing (as in the proof of Theorem 2.6) gives the stated form.

(iv) The greedy selection at round t captures at least $(\eta^{(t)})^2/m^{(t)}$ where $m^{(t)} = m - t$ is the number of remaining private directions. Each round eliminates one, so $\eta^{(m)} = 0$. □

The convergence rate depends on how K-relevant information is distributed across private directions. If it is uniform, $(\eta^{(t)})^2 = (\eta^{(0)})^2(m - t)/m$ and convergence is linear. If it is concentrated in $r < m$ directions, convergence occurs in r rounds. In general, the greedy bound gives $(\eta^{(t)})^2 \leq (\eta^{(0)})^2 \cdot (m - t)/m$.

Adversarial dynamics.

Under adversarial dynamics, models may strategically withhold information. Define the adversarial revelation rate

$$\gamma^{(t)} = \frac{\langle w, d_B^{(t)} \rangle^2}{\max_{\substack{d \in V_B^{(t)},\, d \perp V_A^{(t)} \\ \|d\|=1}} \langle w, d \rangle^2} \in [0, 1],$$

where $\gamma^{(t)} = 1$ is honest and $\gamma^{(t)} = 0$ reveals only K-irrelevant directions.

Proposition 19 (Adversarial Slowdown)**.** *Under adversarial dynamics with* $\gamma^{(t)} \geq \gamma > 0$*:*

(i) $\eta^{(t)} \to 0$*, but convergence requires* $O(m/\gamma)$ *rounds.*

(ii) If $\gamma = 0$ *(complete withholding),* $\eta^{(t)} = \eta^{(0)}$ *for all* t *and no knowledge transfer occurs.*

Proof. (i) The adversarial model reveals a direction capturing at least fraction γ of the greedy optimum, so $\langle w, d_B^{(t)} \rangle^2 \geq \gamma \cdot (\eta^{(t)})^2/(m-t)$. By (24), $(\eta^{(t)})^2$ has multiplicative decay $(1-\gamma/(m-t))$ per round, converging in $O(m/\gamma)$ rounds. (ii) If $d_B^{(t)} \perp w$ at every round, (24) gives no decrease. □

Remark 5. *The revelation rate* γ *connects to the adversarial incentive parameter* λ *from Proposition 3.4. Below* λ^**, debaters have incentive to reveal honestly (*$\gamma \approx 1$*) and convergence is fast. Above* λ^**, defection dominates and* $\gamma \to 0$*, stalling knowledge transfer. The dynamic model thus operationalizes the static coordination failure: the phase transition in* λ *corresponds to a discontinuous drop in* γ*, turning finite-round convergence into non-convergence.*

Remark 6. *Propositions 18 and 19 together predict that the number of useful debate rounds is bounded by the effective rank of private knowledge: the number of private directions carrying substantial* K*-relevant information. For fine-tuned variants of a shared base model, this effective rank is likely small (the fine-tuning perturbation is low-dimensional), predicting that short debates suffice.*

For models with different pretraining corpora, the effective rank could be large, predicting that longer debates are proportionally more valuable.

4 Connection to Constitutional Ambiguity

The debate advantage interacts with the quality of the constitutional specification in a way that connects our work to computational hardness results for value specification.

Definition 20 (Constitutional Ambiguity)**.** *The ε-ambiguity set of K is $\mathcal{A}(\varepsilon) = \{y \in \mathcal{Y} : K(y) \geq K^* - \varepsilon\}$. The constitution is unambiguous if $|\mathcal{A}(0)| = 1$ and ambiguous if $|\mathcal{A}(0)| > 1$.*

Proposition 21 (Ambiguity–Divergence Interaction)**.** *Under the same-corpus assumption ($\eta = 0$):*

1. *If K is unambiguous, debate and RLAIF select the same unique output.*
2. *If K is ambiguous, both select from $\mathcal{A}(0)$ weighted by the base distribution P:*

$$\lim_{\beta \to \infty} \pi^{\beta}_{\text{RLAIF}}(y) = \frac{P(y) \cdot \mathbf{1}[y \in \mathcal{A}(0)]}{\sum_{y' \in \mathcal{A}(0)} P(y')}. \tag{26}$$

Debate converges to the same distribution since debaters have no basis for preferring one K-optimal output over another beyond their shared prior P.

Under knowledge divergence ($\eta > 0$), debate may resolve ambiguities that RLAIF cannot by surfacing private knowledge that distinguishes between elements of $\mathcal{A}(0)$.

The practical import is that debate's value has two independent sources in knowledge divergence (Theorem 6) and ambiguity resolution. Detecting whether a constitution is ambiguous, namely whether $|\mathcal{A}(0)| > 1$, is NP-hard, as it reduces to finding interpretive loopholes in natural-language specifications (Young, 2026). This means we cannot efficiently determine in advance whether debate would help, even if they could measure knowledge divergence.

5 Discussion

One-sided revelation (Proposition 15) is a protocol for eliciting latent knowledge (ELK) (Christiano et al., 2021). Standard ELK asks how to extract knowledge from a model's internal representations. Our work reframes this. If a second model possesses independent knowledge, adversarial interaction can force the first model to externalize private information. The "probe" in this formulation is not an interpretability tool but a model with complementary training data. The private information value η quantifies exactly how much latent knowledge is elicitable through this mechanism.

This is, to our knowledge, the first theory work where the content of models' training data determines whether an oversight protocol is effective. Existing debate theory treats provers as abstract agents; existing RLAIF theory takes the pretrained model as given. We bridges these abstractions by positing the geometry of representation subspaces is a downstream consequence of corpus composition, and it governs whether debate adds value over simpler methods.

Irving et al. (2018) notes that "symmetry between the agents' capabilities is easy to achieve, since we can use the same weights for both agents via self play." Our work reveals this as a feature that eliminates debate's advantage because same-weight debate corresponds to $\theta_i = 0$ which is the degenerate case. The complexity-theoretic results on debate (Brown-Cohen et al., 2024, 2025, 2026) establish what debate can solve in principle, but our framework shows these results have practical bite only in the knowledge-divergent regime that no current work studies empirically.

Goel et al. (2025) empirically demonstrate that as frontier models become more capable, their errors become more correlated, and that diversity between models improves oversight quality. Our framework provides a theoretical explanation as models converge on the same training data and representations, principal angles shrink and the debate advantage vanishes. The empirical benefits of model diversity in oversight are consistent with the $\tan(\theta/2)$ scaling.

Full corpus disjointness is the theoretically clean extreme, but real-world knowledge divergence arises from multiple sources like different fine-tuning data, different model capacities inducing different triage of tail knowledge, and different training-run randomness. Each of these produces nonzero principal angles; small for high-frequency shared knowledge, potentially large for specialized domains. Models fine-tuned on different specializations (medical, legal, scientific corpora) provide the most natural setting for knowledge-divergent debate, with the shared base model corresponding to near-zero angles in general knowledge and large angles in domain-specific directions.

Limitations

The assumption $K(y) = \langle w, h(y) \rangle$ is a simplification. Real constitutional scoring functions are nonlinear; a constitution such as "be helpful, harmless, and honest" involves logical conjunctions over multiple criteria, thresholding effects (an output is harmful or it is not), and context-dependent weighting that resists representation as a single inner product. Our work inherits the linear assumption from the representation geometry literature (Park et al., 2024; Arditi et al., 2024; Rimsky et al., 2024; Turner et al., 2024), where it has proven empirically productive despite its idealization.

The assumption is most defensible in a local regime. Near a particular optimum, a smooth scoring function admits a first-order Taylor expansion that is linear in the representation, and the principal angle structure governs local debate advantage. The global geometry may differ. For instance, a constitutional function with multiple disjoint regions of high score (corresponding to qualitatively different "good" outputs) would not be captured by a single preference direction w. A multi-objective extension replacing w with a set of preference directions $\{w_1, \ldots, w_m\}$ would address this, with debate advantage defined per objective and aggregated. We leave such extensions to future work, noting that the key qualitative insight that debate advantage depends on the angular relationship between models' subspaces and the preference structure does not depend on linearity.

We relax the static subspace assumption by modeling cooperative and adversarial dynamics. However, the dynamic model assumes a stylized absorption process (one direction per round, greedy selection). In practice, in-context learning during debate is a more complex phenomenon: a model reading its opponent's argument may form new associations or activate latent knowledge through mechanisms that do not reduce to the absorption of a single direction. The dynamic framework captures the qualitative distinction between cooperative expansion and adversarial narrowing of subspaces, but the quantitative convergence rates are upper bounds on the idealized process rather than predictions about real debate dynamics.

We assume debate reaches Nash equilibrium, with both debaters playing optimal strategies. In practice, trained debaters may not find equilibrium strategies due to optimization challenges, limited training signal, or the difficulty of learning optimal play in complex sequential games. The debate advantage Δ as defined in our framework describes the ceiling of debate's value, not its floor.

The gap between theoretical and realized debate advantage is itself an interesting quantity. If trained debaters consistently fail to realize the predicted advantage in the knowledge-divergent regime, this would suggest that the game-theoretic structure of debate is harder to learn than the complexity-theoretic results imply. Conversely, if debaters exceed our bounds (which is possible if in-context learning expands their effective subspaces, as discussed above), this would indicate that the static-knowledge assumption is the binding constraint. The dynamic analysis partially addresses this as the convergence rate under adversarial dynamics provides a more realistic bound than the static equilibrium analysis, since it accounts for strategic withholding. If trained debaters exceed our static bounds, this likely indicates that in-context learning is expanding their effective subspaces beyond what the static framework captures.

Our framework also assumes that the judge perfectly evaluates according to K. In reality, human judges are imperfect and potentially manipulable (Khan et al., 2024). Judge imperfection introduces noise into the evaluation, which may differentially affect the three regimes. In the shared-knowledge regime, noise is irrelevant (both protocols converge anyway); in the one-sided regime, a noisy judge may fail to recognize the privately-known better output; in the compositional regime, judge noise could prevent recognition of the composed output even when debaters successfully construct it. A fuller analysis would parameterize judge quality alongside knowledge divergence.

We assume a shared representation map $h : \mathcal{Y} \to \mathbb{R}^d$ that embeds both models' outputs into a common space. This is the assumption that makes principal angles between V_A and V_B well-defined. If models use fundamentally incommensurable representations such as different tokenizations, different embedding dimensions, different representational structures, then a preliminary alignment step is required before the framework applies.

Methods for representation alignment exist. Centered kernel alignment and linear stitching have been used to compare representations across models. However, these methods introduce their own approximation errors, and the resulting principal angles are artifacts of the alignment procedure as much as properties of the underlying models.

The degree to which principal angles are intrinsic (model-dependent) versus extrinsic (alignment-procedure-dependent) is an open question.

For models sharing a common architecture and pretraining pipeline but differing in fine-tuning data, the shared-h assumption is most natural as the base model provides a common representational frame, and fine-tuning perturbs the subspaces within that frame. This is the setting where our framework is most directly applicable, and conveniently also the setting of greatest practical relevance.

References

Andy Arditi, Oscar Balcells Obeso, Aaquib Syed, Daniel Paleka, Nina Rimsky, Wes Gurnee, and Neel Nanda. 2024. Refusal in language models is mediated by a single direction. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, and 32 others. 2022. Constitutional ai: Harmlessness from ai feedback. *Preprint*, arXiv:2212.08073.

Jonah Brown-Cohen, Geoffrey Irving, Simon C. Marshall, Ilan Newman, Georgios Piliouras, and Mario Szegedy. 2026. Debate is efficient with your time. *Preprint*, arXiv:2602.08630.

Jonah Brown-Cohen, Geoffrey Irving, and Georgios Piliouras. 2024. Scalable AI safety via doubly-efficient debate. In *Forty-first International Conference on Machine Learning*.

Jonah Brown-Cohen, Geoffrey Irving, and Georgios Piliouras. 2025. Avoiding obfuscation with prover-estimator debate. *Preprint*, arXiv:2506.13609.

Yongqiang Chen, Gang Niu, James Cheng, Bo Han, and Masashi Sugiyama. 2025. Towards scalable oversight with collaborative multi-agent debate in error detection. *Preprint*, arXiv:2510.20963.

Paul Christiano, Ajeya Cotra, and Mark Xu. 2021. Eliciting latent knowledge: How to tell if your eyes deceive you. Alignment Research Center.

Shashwat Goel, Joschka Strüber, Ilze Amanda Auzina, Karuna K Chandra, Ponnurangam Kumaraguru, Douwe Kiela, Ameya Prabhu, Matthias Bethge, and Jonas Geiping. 2025. Great models think alike and this undermines AI oversight. In *Forty-second International Conference on Machine Learning*.

Geoffrey Irving, Paul Christiano, and Dario Amodei. 2018. Ai safety via debate. *Preprint*, arXiv:1805.00899.

Akbir Khan, John Hughes, Dan Valentine, Laura Ruis, Kshitij Sachan, Ansh Radhakrishnan, Edward Grefenstette, Samuel R. Bowman, Tim Rocktäschel, and Ethan Perez. 2024. Debating with more persuasive llms leads to more truthful answers. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Kiho Park, Yo Joong Choe, and Victor Veitch. 2024. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. 2024. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522, Bangkok, Thailand. Association for Computational Linguistics.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. 2024. Steering language models with activation engineering. *Preprint*, arXiv:2308.10248.

Robin Young. 2026. Np-hard lower bound complexity for semantic self-verification. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.