The Synthetic Consensus Problem in Multi-Agent Knowledge Systems

ZiodbergResearch

February 2026

Abstract

Multi-agent AI systems increasingly employ consensus mechanisms—such as majority voting, debate protocols, and aggregation frameworks—under the assumption that agreement among independent agents constitutes evidence of correctness. We identify and formally define the *synthetic consensus problem*: a failure mode in which multiple AI agents converge on shared conclusions not through independent reasoning over evidence, but through shared training data, architectural homogeneity, and reinforcement learning from human feedback (RLHF) alignment pressure. We distinguish synthetic consensus from genuine independent agreement by developing an information-theoretic framework grounded in mutual information decomposition. We analyze three primary mechanisms that produce synthetic consensus—training data overlap, architectural monoculture, and RLHF-induced mode collapse—and derive bounds on the effective independence of agents under each mechanism. We demonstrate that synthetic consensus can systematically undermine ensemble reliability, producing confident but correlated failure modes that evade detection by standard aggregation methods. Finally, we propose a suite of mitigations including architectural diversity mandates, adversarial agent injection, and provenance tracking systems, and we evaluate their theoretical effectiveness. Our analysis has direct implications for the design of multi-agent systems used in high-stakes domains including scientific review, medical diagnosis, and safety-critical decision making.

1 Introduction

The deployment of multi-agent AI systems has accelerated dramatically in recent years, driven by the intuition that aggregating outputs from multiple models can improve accuracy, robustness, and calibration [Wang et al., 2023a, Du et al., 2023, Liang et al., 2023]. This intuition draws on deep results from statistics and social choice theory: Condorcet's jury theorem guarantees that majority voting among independent agents, each with accuracy exceeding 0.5, converges to perfect accuracy as the number of agents grows [Condorcet, 1785]. The "wisdom of crowds" phenomenon further supports the idea that aggregation reduces noise and surfaces signal [Surowiecki, 2005].

However, these guarantees rest on a critical assumption: *independence*. Condorcet's theorem fails—and ensemble performance can actually degrade—when agents' errors are positively correlated [Ladha, 1992, Dietrich and List, 2008]. In the context of modern large language models (LLMs), there are strong reasons to believe that the independence assumption is systematically violated. Contemporary LLMs share training data sourced from overlapping internet corpora, employ nearly identical transformer architectures, and undergo alignment procedures that push their outputs toward similar distributions over response space.

We term the resulting phenomenon *synthetic consensus*: the appearance of independent agreement that in fact arises from shared inductive biases, common training signals, and convergent optimization pressures. Synthetic consensus is particularly insidious because it mimics the surface-level properties of genuine agreement—multiple agents producing the same answer—while lacking the epistemically valuable property of independent corroboration.

This paper makes the following contributions:

1. We provide a formal definition of synthetic consensus using an information-theoretic framework that decomposes inter-agent agreement into genuine and spurious components (Section 2).

2. We distinguish synthetic consensus from genuine independent agreement and characterize the conditions under which each arises (Section 3).

3. We analyze three primary mechanisms that produce synthetic consensus: training data overlap, architectural homogeneity, and RLHF align-

ment pressure (Section 4).

4. We derive theoretical results on the degradation of ensemble reliability under synthetic consensus (Section 5).

5. We propose and evaluate mitigations including architectural diversity, adversarial agents, and provenance tracking (Section 6).

2 Formal Definition of Synthetic Consensus

We begin by establishing a formal framework for reasoning about consensus in multi-agent systems. Let $\mathcal{A} = \{A_1, A_2, \ldots, A_n\}$ be a set of n AI agents, and let $q \in \mathcal{Q}$ be a query drawn from some space of questions. Each agent A_i produces a response $R_i = A_i(q)$ drawn from a response space $\mathcal{R}$. Let $R^* \in \mathcal{R}$ denote the ground truth or ideal response.

Definition 1 (Consensus)**.** *A set of agents $\mathcal{A}$ exhibits* consensus *on query q if there exists a response $r \in \mathcal{R}$ such that $R_i = r$ for all $i \in \{1, \ldots, n\}$, or more generally, if $d(R_i, R_j) < \epsilon$ for all pairs (i, j) under some distance metric $d : \mathcal{R} \times \mathcal{R} \to \mathbb{R}_{\geq 0}$ and threshold $\epsilon > 0$.*

To distinguish genuine from synthetic consensus, we decompose the mutual information between any two agents' responses.

Definition 2 (Response Mutual Information Decomposition)**.** *For agents A_i and A_j, the mutual information between their responses can be decomposed as:*

$$I(R_i; R_j) = I(R_i; R_j|\mathcal{E}) + I(R_i; R_j; \mathcal{E}) \tag{1}$$

where $\mathcal{E}$ represents the evidence relevant to query q, $I(R_i; R_j|\mathcal{E})$ is the residual mutual information *(agreement not explained by shared evidence), and $I(R_i; R_j; \mathcal{E})$ is the* evidence-mediated mutual information *(agreement attributable to shared evidence).*

The residual mutual information can be further decomposed by identifying its sources:

$$I(R_i; R_j|\mathcal{E}) = I_{\text{data}}(R_i; R_j|\mathcal{E}) + I_{\text{arch}}(R_i; R_j|\mathcal{E}) + I_{\text{align}}(R_i; R_j|\mathcal{E}) + I_{\text{other}}(R_i; R_j|\mathcal{E}) \tag{2}$$

where I_{data}, I_{arch}, and I_{align} represent contributions from shared training data, architectural similarity, and alignment pressure, respectively.

Definition 3 (Synthetic Consensus)**.** *A consensus among agents $\mathcal{A}$ on query q is* synthetic *to degree σ if:*

$$\sigma(q, \mathcal{A}) = \frac{\mathbb{E}_{i \neq j}\left[I(R_i; R_j | \mathcal{E})\right]}{\mathbb{E}_{i \neq j}\left[I(R_i; R_j)\right]} \tag{3}$$

where $\sigma \in [0, 1]$. A consensus is fully synthetic *when $\sigma = 1$ (all agreement is residual) and* fully genuine *when $\sigma = 0$ (all agreement is evidence-mediated).*

Definition 4 (Effective Independence Number)**.** *The* effective independence number *of an ensemble $\mathcal{A}$ with respect to query distribution $\mathcal{Q}$ is:*

$$n_{\mathit{eff}}(\mathcal{A}, \mathcal{Q}) = \frac{\left(\sum_{i=1}^{n} \sigma_i\right)^2}{\sum_{i=1}^{n} \sigma_i^2 + 2\sum_{i<j} \mathit{Cov}(\epsilon_i, \epsilon_j)} \tag{4}$$

where $\sigma_i^2 = \mathit{Var}(\epsilon_i)$ is the variance of agent A_i's error $\epsilon_i = R_i - R^*$, and $\mathit{Cov}(\epsilon_i, \epsilon_j)$ is the covariance between agents' errors. When errors are independent, $n_{\mathit{eff}} = n$. Under perfect positive correlation, $n_{\mathit{eff}} = 1$.

3 Distinguishing Synthetic from Genuine Consensus

Genuine independent agreement and synthetic consensus are observationally equivalent at the output level: both produce n agents asserting the same conclusion. The distinction lies in the causal structure underlying the agreement.

Definition 5 (Genuine Independent Agreement)**.** *Consensus among $\mathcal{A}$ is* genuinely independent *if each agent's response is conditionally independent of all other agents' responses given the evidence:*

$$R_i \perp\!\!\!\perp R_j \mid \mathcal{E}, \quad \forall i \neq j \tag{5}$$

Under genuine independence, the causal graph takes the form of a collider: $R_i \leftarrow \mathcal{E} \rightarrow R_j$, with no other paths connecting R_i and R_j. Under synthetic consensus, additional paths exist through shared confounders:

$$R_i \leftarrow \mathcal{D}_{\text{train}} \rightarrow R_j, \quad R_i \leftarrow \mathcal{F}_{\text{arch}} \rightarrow R_j, \quad R_i \leftarrow \mathcal{L}_{\text{RLHF}} \rightarrow R_j \tag{6}$$

where $\mathcal{D}_{\text{train}}$ represents shared training data, $\mathcal{F}_{\text{arch}}$ represents architectural similarity, and $\mathcal{L}_{\text{RLHF}}$ represents shared alignment objectives.

Proposition 1 (Diagnostic Criterion)**.** *Let $\mathcal{Q}_{novel}$ be a set of queries for which no relevant information exists in any agent's training data. If agents exhibit consensus on $\mathcal{Q}_{novel}$ at a rate significantly above chance, this constitutes evidence of synthetic consensus:*

$$P\left(\textit{consensus} \mid q \in \mathcal{Q}_{novel}\right) \gg P\left(\textit{consensus} \mid q \in \mathcal{Q}_{novel}, \textit{independence}\right) \quad (7)$$

This diagnostic exploits the fact that genuine agreement requires shared evidence, while synthetic agreement persists even in the absence of relevant evidence. Empirically, this can be tested by presenting agents with novel or fabricated questions and measuring agreement rates.

3.1 The Epistemic Value of Consensus

The epistemic value of consensus—the degree to which agreement should update our beliefs toward correctness—is directly related to the degree of independence. By a Bayesian analysis, the posterior probability of a proposition p given unanimous agreement among n agents is:

$$P(p \mid R_1 = \cdots = R_n = p) = \frac{P(p) \cdot P(R_1 = \cdots = R_n = p \mid p)}{P(R_1 = \cdots = R_n = p)} \quad (8)$$

Under full independence with individual accuracy α:

$$P(p \mid \text{unanimous}) = \frac{P(p) \cdot \alpha^n}{P(p) \cdot \alpha^n + (1 - P(p)) \cdot (1 - \alpha)^n} \quad (9)$$

Under perfect correlation (fully synthetic consensus, $n_{\text{eff}} = 1$):

$$P(p \mid \text{unanimous, synthetic}) = \frac{P(p) \cdot \alpha}{P(p) \cdot \alpha + (1 - P(p)) \cdot (1 - \alpha)} \quad (10)$$

The ratio of evidence strength is therefore $\left(\frac{\alpha}{1-\alpha}\right)^{n-1}$, which can be enormous for large n. This quantifies the epistemic loss from synthetic consensus: n correlated agents provide no more evidence than a single agent.

4 Mechanisms of Synthetic Consensus

We now analyze the three primary mechanisms that produce synthetic consensus in contemporary multi-agent AI systems.

4.1 Training Data Overlap

Modern LLMs are trained on massive internet corpora that, despite variations in curation, draw from a largely shared pool of source material. Let $\mathcal{D}_i$ denote the training corpus for agent A_i. The *data overlap coefficient* between agents A_i and A_j is:

$$\omega_{ij} = \frac{|\mathcal{D}_i \cap \mathcal{D}_j|}{|\mathcal{D}_i \cup \mathcal{D}_j|} \tag{11}$$

For contemporary LLMs, empirical estimates suggest $\omega_{ij} \geq 0.7$ for most pairs of frontier models, as Common Crawl, Wikipedia, published books, and curated web text form the backbone of virtually all large-scale training corpora [Dodge et al., 2021, Gao et al., 2020].

Theorem 1 (Data Overlap Correlation Bound)**.** *Let agents A_i and A_j be trained on corpora with overlap coefficient ω_{ij}, and assume that each agent's response to query q is determined by the subset of training data relevant to q. Then the correlation between their error patterns satisfies:*

$$\rho(\epsilon_i, \epsilon_j) \geq \omega_{ij}^{\beta} \tag{12}$$

where $\beta > 0$ depends on the learning algorithm's sensitivity to data composition and typically satisfies $\beta \in [0.5, 2.0]$ for transformer-based models.

Proof sketch. Consider the partition of each agent's training data into shared ($\mathcal{D}_i \cap \mathcal{D}_j$) and unique ($\mathcal{D}_i \setminus \mathcal{D}_j$, $\mathcal{D}_j \setminus \mathcal{D}_i$) components. The agent's learned representation can be decomposed as $f_i = f_{\text{shared}} + f_{\text{unique},i}$ where f_{shared} depends only on $\mathcal{D}_i \cap \mathcal{D}_j$. Under standard concentration assumptions, the contribution of f_{shared} to the agent's output scales as ω_{ij}^{β} relative to total output variance. Since f_{shared} is identical for both agents, it induces correlated errors with correlation at least ω_{ij}^{β}. □

The training data overlap problem is compounded by *data provenance convergence*: as the internet becomes increasingly populated by AI-generated content, future training corpora will contain outputs from previous-generation models, creating feedback loops that amplify consensus-producing signals [Shumailov et al., 2023].

4.2 Architectural Homogeneity

The overwhelming dominance of the transformer architecture [Vaswani et al., 2017] in modern LLMs constitutes an architectural monoculture. We formalize the effect of architectural similarity on consensus.

Definition 6 (Architectural Inductive Bias)**.** *The* inductive bias *of architecture $\mathcal{F}$ is the function $b_{\mathcal{F}} : \mathcal{Q} \to \Delta(\mathcal{R})$ that maps queries to response distributions in the absence of training data (or equivalently, the implicit prior over functions induced by the architecture and random initialization).*

When two agents share the same architecture, $b_{\mathcal{F}_i} = b_{\mathcal{F}_j}$, and their inductive biases are perfectly aligned. This manifests in several concrete ways:

Attention pattern convergence. Transformer models tend to develop similar attention patterns for similar inputs, a phenomenon documented as "attention head universality" [Olsson et al., 2022]. Specifically, induction heads, which implement in-context copying behavior, emerge reliably across independent training runs on different data, suggesting that the transformer architecture deterministically channels learning toward certain representational strategies.

Representational alignment. Studies on representation similarity between independently trained models have found that neural networks with the same architecture converge to similar internal representations up to linear transformation [Li et al., 2015, Kornblith et al., 2019]. Formally, if ϕ_i and ϕ_j are the learned representations of two transformers, there often exists a linear map W such that $\phi_j \approx W\phi_i$, implying that both models organize knowledge similarly and will therefore make correlated errors.

Expressivity constraints. The set of functions efficiently expressible by a transformer of given depth L and width d is constrained. Certain functions lie outside this set regardless of training data, creating systematic shared blind spots. Let $\mathcal{H}_{\mathcal{F}}$ denote the hypothesis class of architecture $\mathcal{F}$. Then for any query q whose ideal response R^* requires a function $f \notin \mathcal{H}_{\mathcal{F}}$, all agents using architecture $\mathcal{F}$ will err, producing correlated failures.

Proposition 2 (Architectural Correlation)**.** *For agents sharing architecture $\mathcal{F}$, the minimum error correlation attributable to shared inductive bias satisfies:*

$$\rho_{arch}(\epsilon_i, \epsilon_j) \geq 1 - \frac{dim(\mathcal{H}_{\mathcal{F}}^{\perp} \cap span(\mathcal{Q}))}{dim(span(\mathcal{Q}))} \tag{13}$$

where $\mathcal{H}_{\mathcal{F}}^{\perp}$ is the orthogonal complement of the hypothesis class in function space.

4.3 RLHF Alignment Pressure

Reinforcement Learning from Human Feedback [Christiano et al., 2017, Ouyang et al., 2022] has become the standard method for aligning LLMs with human preferences. We argue that RLHF introduces a powerful consensus-producing force that operates independently of training data and architecture.

The RLHF objective can be written as:

$$\mathcal{L}_{\text{RLHF}}(\theta) = \mathbb{E}_{q\sim\mathcal{Q}, r\sim\pi_\theta(\cdot|q)}\left[R_\phi(q,r)\right] - \beta\,\text{KL}\left[\pi_\theta(\cdot|q)\|\pi_{\text{ref}}(\cdot|q)\right] \tag{14}$$

where R_ϕ is the learned reward model and β controls the KL penalty against the reference policy π_{ref}.

Definition 7 (RLHF Mode Collapse)**.** RLHF mode collapse *occurs when the alignment procedure causes the output distribution to concentrate on a small subset of response space:*

$$H(\pi_{aligned}(\cdot|q)) \ll H(\pi_{base}(\cdot|q)) \tag{15}$$

where $H(\cdot)$ denotes entropy.

RLHF produces synthetic consensus through three sub-mechanisms:

Reward model homogeneity. Reward models are trained on human preference data that reflects systematic biases: preferences for confident-sounding responses, longer and more detailed outputs, hedging language, and particular rhetorical structures [Casper et al., 2023]. Since these biases are consistent across human annotators (reflecting shared cultural and cognitive patterns), reward models trained on different annotator pools converge to similar preference orderings. Let R_{ϕ_i} and R_{ϕ_j} be reward models for different alignment processes. Empirically, rank correlation $\tau(R_{\phi_i}, R_{\phi_j}) > 0.8$ for most frontier model pairs.

Sycophancy and agreement bias. RLHF training incentivizes agents to produce responses that humans rate highly, which creates pressure toward producing responses that match common human beliefs—even when those beliefs are incorrect [Perez et al., 2022, Sharma et al., 2023]. This induces correlation between agents specifically in error cases, which is the most damaging form of correlation for ensemble reliability.

Safety-induced refusal correlation. Alignment procedures include safety training that causes models to refuse certain categories of queries. Since safety taxonomies are largely shared across organizations (reflecting common regulatory pressures and industry norms), aligned models exhibit correlated refusal patterns that constitute a form of synthetic consensus on the meta-question of which queries are answerable.

Theorem 2 (RLHF Correlation Amplification)**.** *Let π_1 and π_2 be policies derived from base models with error correlation ρ_0 after RLHF with reward models having rank correlation τ. The post-alignment error correlation satisfies:*

$$\rho_{post} \geq \rho_0 + (1 - \rho_0) \cdot \tau \cdot \left(1 - \frac{\beta}{\beta + \lambda}\right) \tag{16}$$

where λ is the effective reward signal strength and β is the KL penalty coefficient. As $\beta \to 0$ (unconstrained optimization of the reward), $\rho_{post} \to \rho_0 + (1 - \rho_0)\tau$, approaching perfect correlation as $\tau \to 1$.

5 Implications for Multi-Agent Systems

Synthetic consensus has profound implications for multi-agent system designs that rely on agreement as a signal of correctness.

5.1 Degradation of Ensemble Reliability

Consider a majority-vote ensemble of n agents, each with individual accuracy $\alpha > 0.5$. Under independence, the ensemble accuracy is:

$$P_{\text{correct}}^{\text{indep}} = \sum_{k=\lceil n/2 \rceil}^{n} \binom{n}{k} \alpha^k (1 - \alpha)^{n-k} \tag{17}$$

Under synthetic consensus with effective independence number $n_{\text{eff}} < n$, the ensemble behaves as if it contains only n_{eff} independent agents. For the limiting case of $n_{\text{eff}} = 1$:

$$P_{\text{correct}}^{\text{synthetic}} = \alpha \tag{18}$$

regardless of n. Adding more correlated agents provides no improvement. Worse, if RLHF alignment pressure has introduced systematic bias toward popular-but-incorrect answers on certain query types, the ensemble accuracy can be *below* α due to confidence amplification effects.

5.2 Failure of Debate Protocols

AI debate protocols [Irving et al., 2018, Du et al., 2023] rely on the assumption that competing agents will surface different perspectives and identify errors in each other's reasoning. Under synthetic consensus, debating agents share the same blind spots and are unlikely to challenge conclusions that arise from shared training artifacts.

Proposition 3 (Debate Limitation)**.** *Let agents A_1 and A_2 engage in a debate protocol on query q. If the agents share a common misconception m (a confident incorrect belief arising from shared training data), the probability that the debate surfaces the error is bounded by:*

$$P(\textit{error detected via debate}) \leq 1 - P(m \mid A_1) \cdot P(m \mid A_2) \leq 1 - \omega_{12}^{2\beta} \tag{19}$$

For $\omega_{12} = 0.8$ and $\beta = 1$, this gives $P(\textit{error detected}) \leq 0.36$, meaning shared misconceptions survive debate more than 60% of the time.

5.3 Calibration Illusions

Multi-agent systems often use agreement rates as a calibration signal: high agreement is interpreted as high confidence, and disagreement triggers escalation or abstention. Under synthetic consensus, agreement rates are inflated, producing overconfident predictions. The calibration error induced by synthetic consensus can be quantified as:

$$\text{ECE}_{\text{synthetic}} = \mathbb{E}\left[\left|P(\text{correct} \mid \hat{p}_{\text{agreement}}) - \hat{p}_{\text{agreement}}\right|\right] \geq (1 - 1/n_{\text{eff}}) \cdot \text{Corr}_{\text{systematic}} \tag{20}$$

where $\hat{p}_{\text{agreement}}$ is the agreement-based confidence estimate and $\text{Corr}_{\text{systematic}}$ quantifies the strength of systematic errors.

5.4 Vulnerability to Correlated Adversarial Inputs

Synthetic consensus implies that adversarial examples transfer between agents at elevated rates. An input x' crafted to fool agent A_i will fool agent A_j with probability at least $\rho(\epsilon_i, \epsilon_j)$, compared to the base rate under independence. This has severe implications for safety-critical multi-agent systems that use redundancy for fault tolerance.

6 Proposed Mitigations

We propose a portfolio of mitigations targeting each mechanism of synthetic consensus.

6.1 Architectural Diversity

Definition 8 (Diversity-Constrained Ensemble)**.** *An ensemble $\mathcal{A}$ satisfies the* δ-diversity constraint *if for all pairs (A_i, A_j):*

$$d_{arch}(\mathcal{F}_i, \mathcal{F}_j) \geq \delta \tag{21}$$

where d_{arch} is a metric on architecture space (e.g., based on computational graph edit distance, or divergence of inductive biases on a reference task suite).

Concretely, diverse ensembles should include models with fundamentally different architectures: transformers, state-space models (e.g., Mamba [Gu and Dao, 2023]), mixture-of-experts architectures, retrieval-augmented models, and neuro-symbolic systems. Each architecture class introduces different inductive biases, reducing ρ_{arch}.

Proposition 4 (Diversity Benefit)**.** *An ensemble satisfying the δ-diversity constraint with K distinct architecture families achieves effective independence number:*

$$n_{eff} \geq K \cdot \left(1 - \max_k \omega^{\beta}_{intra,k}\right) \tag{22}$$

where $\omega_{intra,k}$ is the maximum data overlap within architecture family k.

Algorithm 1 Adversarial Consensus Protocol

Require: Query q, agents $\mathcal{A} = \{A_1, \ldots, A_n\}$, adversarial agents $\mathcal{A}_{\text{adv}} = \{A_{\text{adv},1}, \ldots, A_{\text{adv},m}\}$, threshold τ

Collect responses $\{R_i = A_i(q)\}_{i=1}^{n}$
Compute consensus response $r_c = \text{MajorityVote}(\{R_i\})$
Compute agreement rate $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}[R_i = r_c]$
for $j = 1$ to m **do**
    $C_j = A_{\text{adv},j}(q, r_c, \{R_i\})$ {Generate challenge}
    $s_j = \text{EvaluateChallenge}(C_j, q, r_c)$ {Score challenge strength}
end for
$s_{\max} = \max_j s_j$
if $s_{\max} > \tau$ **then**
    Flag consensus as potentially synthetic
    Escalate to extended deliberation or human review
else
    Accept consensus with confidence $\hat{p} \cdot (1 - s_{\max}/\tau)$
end if

6.2 Adversarial Agent Injection

We propose the deliberate inclusion of *adversarial agents* in multi-agent systems—agents specifically trained or prompted to challenge consensus.

Definition 9 (Adversarial Agent)**.** *An* adversarial agent A_{adv} *is one that, given a consensus response* r *from other agents, is incentivized to find the strongest possible counterargument or alternative response* $r' \neq r$*.*

The adversarial agent framework can be formalized as a two-player game. Let $\pi_{\text{consensus}}$ be the consensus policy and π_{adv} be the adversarial policy. The adversarial objective is:

$$\max_{\pi_{\text{adv}}} \mathbb{E}_{q \sim \mathcal{Q}} \left[\mathbb{1}\left[R^* = A_{\text{adv}}(q, r_{\text{consensus}})\right] \cdot \mathbb{1}\left[R^* \neq r_{\text{consensus}}\right] \right] \tag{23}$$

This objective rewards the adversarial agent for correctly identifying cases where the consensus is wrong.

A critical design consideration is ensuring that adversarial agents are themselves not subject to synthetic consensus with the primary agents. This can be achieved by drawing adversarial agents from different architecture

families, training them on different data, or using non-neural approaches (e.g., formal verification systems, symbolic reasoners) as adversarial agents.

6.3 Provenance Tracking

We propose a provenance tracking system that traces the evidential basis for each agent's response, enabling direct measurement of the synthetic consensus degree.

Definition 10 (Response Provenance)**.** *The* provenance *of agent A_i's response R_i to query q is a set $\mathcal{P}_i(q) \subseteq \mathcal{D}_i \cup \{q\}$ identifying the training examples and input features that causally contributed to R_i.*

Given provenance information, the synthetic consensus degree can be directly estimated:

$$\hat{\sigma}(q, \mathcal{A}) = \frac{\sum_{i<j} |\mathcal{P}_i(q) \cap \mathcal{P}_j(q)|}{\sum_{i<j} |\mathcal{P}_i(q) \cup \mathcal{P}_j(q)|} \tag{24}$$

Provenance tracking can be implemented through several mechanisms:

Influence functions. Influence functions [Koh and Liang, 2017] estimate the effect of each training example on a model's prediction, providing an approximation to provenance. For a response R_i to query q, the influence of training example z is:

$$\mathcal{I}(z, q) = -\nabla_\theta \ell(q, R_i)^\top H_\theta^{-1} \nabla_\theta \ell(z, R_z) \tag{25}$$

where H_θ is the Hessian of the training loss. While exact computation is intractable for large models, efficient approximations exist [Grosse et al., 2023].

Retrieval attribution. For retrieval-augmented models, provenance is directly observable through the retrieved documents. Mandating retrieval-augmented architectures in multi-agent systems enables straightforward provenance comparison.

Mechanistic attribution. Advances in mechanistic interpretability [Elhage et al., 2021, Conmy et al., 2023] offer the prospect of identifying which learned circuits contribute to specific outputs, providing a fine-grained provenance signal at the computational rather than data level.

6.4 Training Data Diversification

To reduce ω_{ij}, we propose deliberate diversification of training corpora:

1. **Data partitioning**: Divide available training data into non-overlapping shards and train different models on different shards. While this reduces individual model quality, it substantially increases ensemble n_{eff}.

2. **Temporal stratification**: Train models on data from different time periods, ensuring different factual baselines and reducing agreement from shared factual errors.

3. **Linguistic diversification**: Train models primarily on data from different languages or cultural contexts, as many shared errors in current LLMs reflect English-language and Western-centric biases in training data.

Theorem 3 (Optimal Data Allocation)**.** *Given a total dataset $\mathcal{D}$ and n agents to train, the data allocation strategy that maximizes n_{eff} subject to a minimum individual accuracy constraint $\alpha_{\min}$ is:*

$$|\mathcal{D}_i \cap \mathcal{D}_j| = \max\left(0, \frac{n \cdot |\mathcal{D}_{\min}(\alpha_{\min})| - |\mathcal{D}|}{n-1}\right) \tag{26}$$

where $|\mathcal{D}_{\min}(\alpha_{\min})|$ is the minimum dataset size needed to achieve accuracy $\alpha_{\min}$.

6.5 Alignment Diversification

To counteract RLHF-induced consensus, we propose:

1. **Diverse reward models**: Train reward models on preference data from different annotator populations, cultures, and expertise domains.

2. **Variable KL penalties**: Use different values of the KL penalty β across agents, producing varying degrees of alignment-induced narrowing and preserving greater output diversity.

3. **Alternative alignment methods**: Employ different alignment techniques across the ensemble—RLHF, DPO [Rafailov et al., 2023], constitutional AI [Bai et al., 2022], debate-based alignment—to reduce I_{align}.

7 Related Work

The synthetic consensus problem connects to several established research areas. The study of ensemble diversity has a long history in machine learning [Krogh and Vedelsby, 1994, Kuncheva and Whitaker, 2003]. Hansen and Salamon [1990] showed that ensemble improvement requires diversity among members, and Brown et al. [2005] formalized the ambiguity decomposition linking ensemble error to diversity. Our work extends this analysis to the specific correlations induced by shared training pipelines in LLMs.

In social epistemology, the problem of "informational cascades" [Banerjee, 1992, Bikhchandani et al., 1992] describes how rational agents can converge on incorrect beliefs by following predecessors rather than private signals. Synthetic consensus is an analogous phenomenon at the level of AI systems, where "following predecessors" is replaced by "sharing training data."

The fragility of Condorcet-type results under correlation has been studied by Ladha [1992], Dietrich and List [2008], and Pivato [2017]. Our contribution is to identify the specific mechanisms that produce correlation in LLM ensembles and to quantify their effects.

Recent empirical work has documented surprising agreement patterns among LLMs. Zheng et al. [2023] found high correlation in LLM-as-judge evaluations, and Wang et al. [2023b] observed convergent behavior in multi-agent debate settings. Our framework provides a theoretical explanation for these observations.

8 Limitations and Future Directions

Our analysis has several limitations. First, the exact decomposition of mutual information into data, architecture, and alignment components (Equation 2) is conceptually clean but difficult to estimate empirically, as these factors interact in complex ways. Second, our correlation bounds (Theorems 1 and 2) rely on simplifying assumptions about the separability of different correlation sources. Third, the provenance tracking mechanisms we propose face significant scalability challenges for models with billions of parameters trained on trillions of tokens.

Future work should focus on: (1) empirical measurement of n_{eff} for existing multi-agent systems across diverse task domains; (2) development of practical, scalable provenance tracking systems; (3) investigation of whether

synthetic consensus can be detected purely from output distributions without access to model internals; and (4) extension of the framework to multimodal agents and tool-using agents, where additional sources of correlation (shared tools, APIs, knowledge bases) may exacerbate the problem.

9 Conclusion

We have identified, formally defined, and analyzed the synthetic consensus problem in multi-agent AI systems. Our key insight is that the widespread practice of aggregating outputs from multiple AI agents—whether through voting, debate, or other consensus mechanisms—provides substantially less epistemic value than commonly assumed, because contemporary AI agents are far from independent. The three mechanisms we analyzed—training data overlap, architectural homogeneity, and RLHF alignment pressure—each independently produce significant correlations, and their combined effect can reduce the effective independence number of a nominally large ensemble to near unity.

This finding has immediate practical implications. Multi-agent systems deployed in high-stakes domains should not rely on agreement among architecturally similar, similarly trained, and similarly aligned models as evidence of correctness. Instead, practitioners should actively pursue the mitigations we have outlined: architectural diversity, adversarial agent injection, provenance tracking, and diversification of both training data and alignment procedures.

More broadly, the synthetic consensus problem highlights a fundamental tension in the current AI ecosystem: the pressures that drive convergence—shared training data, proven architectures, standardized alignment procedures—also undermine the diversity that makes multi-agent aggregation valuable. Addressing this tension will require deliberate investment in diverse approaches to AI development, even when individual alternatives may underperform the dominant paradigm.

References

Banerjee, A. V. (1992). A simple model of herd behavior. *The Quarterly Journal of Economics*, 107(3):797–817.

Bai, Y., Kadavath, S., Kundu, S., et al. (2022). Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*.

Bikhchandani, S., Hirshleifer, D., and Welch, I. (1992). A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of Political Economy*, 100(5):992–1026.

Brown, G., Wyatt, J., Harris, R., and Yao, X. (2005). Diversity creation methods: A survey and categorisation. *Information Fusion*, 6(1):5–20.

Casper, S., Davies, X., Shi, C., et al. (2023). Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*.

Christiano, P. F., Leike, J., Brown, T., et al. (2017). Deep reinforcement learning from human preferences. *NeurIPS*, 30.

Condorcet, M. (1785). *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. Paris: Imprimerie Royale.

Conmy, A., Mavor-Parker, A. N., Lynch, A., et al. (2023). Towards automated circuit discovery for mechanistic interpretability. *NeurIPS*, 36.

Dietrich, F. and List, C. (2008). A liberal paradox for judgment aggregation. *Social Choice and Welfare*, 31(1):59–78.

Dodge, J., Sap, M., Marasović, A., et al. (2021). Documenting large webtext corpora: A case study on the Colossal Clean Crawled Corpus. *EMNLP*.

Du, Y., Li, S., Torralba, A., et al. (2023). Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*.

Elhage, N., Nanda, N., Olsson, C., et al. (2021). A mathematical framework for transformer circuits. *Transformer Circuits Thread*.

Gao, L., Biderman, S., Black, S., et al. (2020). The Pile: An 800GB dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Grosse, R., Bae, J., Anil, C., et al. (2023). Studying large language model generalization with influence functions. *arXiv preprint arXiv:2308.03296*.

Gu, A. and Dao, T. (2023). Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.

Hansen, L. K. and Salamon, P. (1990). Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(10):993–1001.

Irving, G., Christiano, P., and Amodei, D. (2018). AI safety via debate. *arXiv preprint arXiv:1805.00899*.

Koh, P. W. and Liang, P. (2017). Understanding black-box predictions via influence functions. *ICML*.

Kornblith, S., Norouzi, M., Lee, H., and Hinton, G. (2019). Similarity of neural network representations revisited. *ICML*.

Krogh, A. and Vedelsby, J. (1994). Neural network ensembles, cross validation, and active learning. *NeurIPS*, 7.

Kuncheva, L. I. and Whitaker, C. J. (2003). Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning*, 51(2):181–207.

Ladha, K. K. (1992). The Condorcet jury theorem, free speech, and correlated votes. *American Journal of Political Science*, 36(3):617–634.

Li, Y., Yosinski, J., Clune, J., et al. (2015). Convergent learning: Do different neural networks learn the same representations? *ICLR*.

Liang, T., He, Z., Jiao, W., et al. (2023). Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.19118*.

Olsson, C., Elhage, N., Nanda, N., et al. (2022). In-context learning and induction heads. *Transformer Circuits Thread*.

Ouyang, L., Wu, J., Jiang, X., et al. (2022). Training language models to follow instructions with human feedback. *NeurIPS*, 35.

Perez, E., Ringer, S., Lukošiūtė, K., et al. (2022). Discovering language model behaviors with model-written evaluations. *arXiv preprint arXiv:2212.09251*.

Pivato, M. (2017). Epistemic democracy with correlated voters. *Journal of Mathematical Economics*, 72:51–69.

Rafailov, R., Sharma, A., Mitchell, E., et al. (2023). Direct preference optimization: Your language model is secretly a reward model. *NeurIPS*, 36.

Sharma, M., Tong, M., Korbak, T., et al. (2023). Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*.

Shumailov, I., Shumaylov, Z., Zhao, Y., et al. (2023). The curse of recursion: Training on generated data makes models forget. *arXiv preprint arXiv:2305.17493*.

Surowiecki, J. (2005). *The Wisdom of Crowds*. Anchor Books.

Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. *NeurIPS*, 30.

Wang, X., Wei, J., Schuurmans, D., et al. (2023). Self-consistency improves chain of thought reasoning in language models. *ICLR*.

Wang, Z., Mao, S., Wu, W., et al. (2023). Unleashing the emergent cognitive synergy in large language models: A task-solving agent through multi-persona self-collaboration. *arXiv preprint arXiv:2307.05300*.

Zheng, L., Chiang, W.-L., Sheng, Y., et al. (2023). Judging LLM-as-a-judge with MT-bench and chatbot arena. *NeurIPS*, 36.