

The Resonant Consensus Protocol

Cross-Cluster Resonance Classification for Multi-Agent Systems

Alan J. Garcia

github.com/garciaalan186

1 Overview

1.1 The Problem: Confident Uncertainty

Large language models face a fundamental tension between helpfulness and epistemic honesty. Standard training objectives prioritize fluent, confident responses, inadvertently incentivizing models to guess rather than acknowledge uncertainty [1, 2]. This produces two failure modes:

1. **Overconfident errors:** The model generates plausible but incorrect content with unwarranted certainty—a phenomenon widely documented as “hallucination” [3, 4].
2. **Sycophantic agreement:** The model adapts responses to align with perceived user preferences, sacrificing accuracy for approval [5, 6].

Both failure modes share a common structure: the model lacks a principled mechanism for representing genuine uncertainty. When multiple valid perspectives exist, or when a query is ambiguous, the model typically selects one framing and commits—leaving the user unaware that the response reflects a contestable position rather than established fact.

1.2 Cross-Cluster Resonance as a Solution

The Resonant Consensus Protocol¹ addresses this challenge by providing an external signal that is harder to game than internal confidence estimates. Rather than asking a single model “how confident are you?”, the protocol solicits evaluations from multiple agents with *adversarial* perspectives and classifies artifacts based on **cross-cluster resonance**.

We define *cross-cluster resonance* as the pattern of endorsement across predefined adversarial groups: an artifact exhibits cross-cluster resonance when approval spans multiple clusters rather than concentrating within a single faction.

The key insight: **agreement between parties with opposing incentives is stronger evidence than agreement between parties with aligned incentives**. An artifact endorsed by both an advocate cluster and a critic cluster has survived genuine cross-examination. An artifact endorsed only by advocates—or rejected by all—carries different epistemic weight.

This classification produces a **Cross-Cluster Resonance State**—a structured summary that tells the orchestrator not just what was said, but how it resonated across perspectives. Crucially, contestation is treated as information, not failure. When artifacts are contested or rejected, the orchestrator gains a principled basis for surfacing uncertainty, requesting clarification, or presenting multiple perspectives—rather than gambling on the “best guess.”

2 Binary Framework ($n = 2$)

We begin with the binary case—two adversarial clusters—which provides the clearest illustration of cross-cluster resonance and forms the foundation for generalization.

¹This work builds on the Networked Survey, a primary research methodology for quantifying consensus developed by the author in 2013.

2.1 Setup

Let C^+ and C^- denote two adversarial clusters (e.g., advocate vs. critic, optimist vs. skeptic). Each cluster evaluates each artifact, producing a binary approval:

$$\text{Approve}_{C^+}(\omega), \text{Approve}_{C^-}(\omega) \in \{0, 1\} \quad (1)$$

This yields four possible outcomes, corresponding to four classification tiers.

2.2 Four-Tier Classification

| | | | |
|--|--------|---------------------------|--------------------|
| | | Approve | |
| | | Negative Polar | Positive Consensus |
| | C^- | | |
| | Reject | Negative Consensus | Positive Polar |
| | | Reject | Approve |
| | | C^+ (Reference Cluster) | |

Positive Consensus — C^+ approves, C^- approves

Both adversarial clusters endorse this artifact. It has survived cross-examination from opposing perspectives and represents robust, cross-cutting resonance.

Negative Consensus — C^+ rejects, C^- rejects

Both adversarial clusters reject this artifact. Cross-cluster agreement that the artifact is flawed, unsupported, or inappropriate.

Positive Polar — C^+ approves, C^- rejects

The reference cluster endorses this artifact, but the opposing cluster rejects it. The artifact reflects the C^+ perspective specifically.

Negative Polar — C^+ rejects, C^- approves

The reference cluster rejects this artifact, but the opposing cluster endorses it. The artifact reflects the C^- perspective specifically.

2.3 Multi-Dimensional Assessment

Rather than prescribing specific actions for each tier, the protocol provides a multi-dimensional assessment that the orchestrator interprets based on context:

| Tier | Contestation | Bias Direction | Risk if Acted Upon |
|--------------------|--------------|----------------|--------------------|
| Positive Consensus | Low | Neutral | Low |
| Negative Consensus | Low | Neutral | High |
| Positive Polar | High | C^+ | Moderate |
| Negative Polar | High | C^- | Moderate |

- **Contestation:** How disputed is this artifact across clusters?
- **Bias Direction:** Which cluster’s perspective does this artifact reflect?
- **Risk if Acted Upon:** Epistemic risk of treating this artifact as authoritative.

This structure enables the orchestrator to make context-appropriate decisions: a Negative Consensus artifact might still be presented to the user with appropriate framing (“both perspectives rejected this approach because...”), while a Positive Polar artifact might prompt the orchestrator to surface the disagreement or request clarification.

3 Generalization ($n > 2$)

The binary framework extends naturally to n clusters by designating a **reference cluster** C_1 and computing a **resonance ratio** across all clusters.

3.1 Components

System Elements

- **Agents** $\mathcal{A} = \{a_1, \dots, a_m\}$ — LLM instances with distinct system prompts
- **Clusters** $\mathcal{C} = \{C_1, \dots, C_n\}$ — Partition of agents into n adversarial groups
- **Reference Cluster** C_1 — Designated cluster for determining positive/negative polarity
- **Artifacts** $\Omega = \{\omega_1, \dots, \omega_k\}$ — Responses generated by agents
- **Votes** $v(\omega, a) \in \{0, 1\}$ — Agent a ’s approval of artifact ω

3.2 Cluster Approval

For each artifact ω and cluster C_i , we determine whether the cluster approves:

$$\text{Approve}_i(\omega) = \begin{cases} 1 & \text{if } \frac{1}{|C_i|} \sum_{a \in C_i} v(\omega, a) \geq \theta_i \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

Where $\theta_i \in (0, 1]$ is the approval threshold for cluster C_i . By default, $\theta_i = \theta$ for all clusters (a global threshold, typically 0.5 for simple majority).

3.3 Approval Set and Resonance Ratio

The **approval set** captures which clusters approved:

$$S(\omega) = \{C_i \in \mathcal{C} : \text{Approve}_i(\omega) = 1\} \quad (3)$$

The **resonance ratio** normalizes this (where $|\cdot|$ denotes the number of elements in a set):

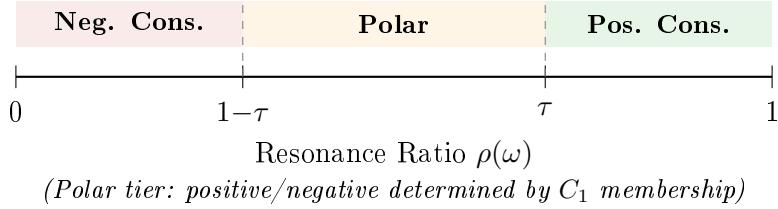
$$\rho(\omega) = \frac{|S(\omega)|}{n} \in [0, 1] \quad (4)$$

3.4 Four-Tier Classification (General Case)

Classification depends on both the resonance ratio $\rho(\omega)$ and whether the reference cluster C_1 approves:

$$\text{Tier}(\omega) = \begin{cases} \textbf{Positive Consensus} & \text{if } \rho(\omega) \geq \tau \text{ and } C_1 \in S(\omega) \\ \textbf{Negative Consensus} & \text{if } \rho(\omega) \leq 1 - \tau \\ \textbf{Positive Polar} & \text{if } 1 - \tau < \rho(\omega) < \tau \text{ and } C_1 \in S(\omega) \\ \textbf{Negative Polar} & \text{if } 1 - \tau < \rho(\omega) < \tau \text{ and } C_1 \notin S(\omega) \end{cases} \quad (5)$$

Where $\tau \in (0.5, 1]$ is the consensus threshold.



3.5 Note on Empty Tiers

For small n or certain values of τ , the Polar tier may be empty (no discrete ρ values fall in the range). This is expected behavior—the system operates in binary mode for that configuration.

4 The Binary Case ($n = 2$)

The most common configuration uses two adversarial clusters. This section illustrates the framework with $n = 2$.

4.1 Setup

Let $\mathcal{C} = \{C^+, C^-\}$ represent two opposing perspectives (e.g., advocate and critic). The approval set $S(\omega)$ can take four values:

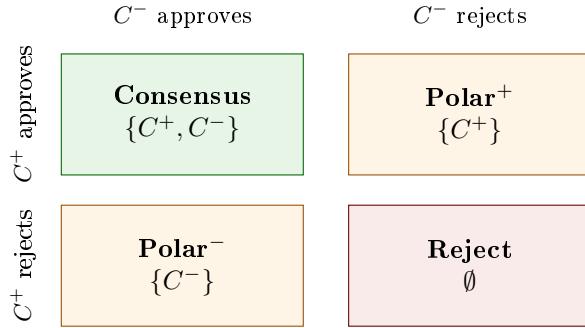
| $S(\omega)$ | $ S $ | ρ | Interpretation |
|----------------|-------|--------|------------------------|
| $\{C^+, C^-\}$ | 2 | 1.0 | Both clusters approve |
| $\{C^+\}$ | 1 | 0.5 | Only advocates approve |
| $\{C^-\}$ | 1 | 0.5 | Only critics approve |
| \emptyset | 0 | 0.0 | Neither approves |

4.2 Classification with $\tau = 0.6$

| $S(\omega)$ | ρ | Tier | Action |
|----------------|--------|-----------|------------------------------------|
| $\{C^+, C^-\}$ | 1.0 | Consensus | GROUND |
| $\{C^+\}$ | 0.5 | Polar | CONTEXTUALIZE as advocate position |
| $\{C^-\}$ | 0.5 | Polar | CONTEXTUALIZE as critic position |
| \emptyset | 0.0 | Reject | EXCLUDE |

4.3 The 2×2 Quadrant View

For $n = 2$, the classification can be visualized as a quadrant:



5 Scoring

Within each tier, artifacts are ranked by **total approval rate**:

$$\text{Score}(\omega) = \frac{1}{|\mathcal{A}| - 1} \sum_{a \neq \text{author}(\omega)} v(\omega, a) \quad (6)$$

This is the fraction of non-authoring agents who approved, regardless of cluster. Higher scores indicate stronger overall support.

5.1 Optional: Consensus Balance Bonus

For Consensus artifacts, balanced approval across clusters may be preferred. An optional refinement penalizes lopsided approval:

$$\text{Score}_{\text{balanced}}(\omega) = \text{Score}(\omega) \times (1 - \sigma_S(\omega)) \quad (7)$$

Where $\sigma_S(\omega)$ is the standard deviation of per-cluster approval rates:

$$\sigma_S(\omega) = \sqrt{\frac{1}{n} \sum_{i=1}^n (r_i(\omega) - \bar{r}(\omega))^2} \quad (8)$$

And $r_i(\omega) = \frac{1}{|C_i|} \sum_{a \in C_i} v(\omega, a)$ is the approval rate within cluster C_i .

6 Persuasive Artifacts

A valuable subclass of Consensus artifacts demonstrate **cross-cluster persuasion**: they were authored by one cluster but approved by adversarial clusters.

Cross-Cluster Persuasion

For a Consensus artifact ω authored by an agent in cluster C_i :

- The artifact is **persuasive** if $C_j \in S(\omega)$ for some $j \neq i$
- The **persuasion reach** is $|S(\omega) \setminus \{C_i\}|$ — how many adversarial clusters it convinced

For $n = 2$, this simplifies to:

- **Accelerator**: Consensus artifact authored by C^+ , approved by C^-
“An advocate’s argument that even critics accept.”

- **Mitigator:** Consensus artifact authored by C^- , approved by C^+
“A critic’s argument that even advocates acknowledge.”

These are high-value artifacts for constructing balanced responses.

7 The Cross-Cluster Resonance State

The protocol outputs a **Cross-Cluster Resonance State** for each artifact—a structured summary providing multi-dimensional assessment for the orchestrator:

Resonance State Object

```
ResonanceState() = {
    artifact: ,
    approval_set: S(),
    resonance_ratio: (),
    tier: PositiveConsensus | NegativeConsensus |
        PositivePolar | NegativePolar,
    // Multi-dimensional assessment
    contestation: Low | High,
    bias_direction: Neutral | C_1 | "non-C_1",
    risk_if_acted_upon: Low | Moderate | High,
    // Metadata
    score: Score(),
    author_cluster: C_i,
    is_persuasive: boolean,
    full_consensus: boolean // true if |S| = n
}
```

7.1 Assessment Dimensions

- **Contestation:** Derived from tier. Consensus tiers (positive or negative) have low contestation; Polar tiers have high contestation.
- **Bias Direction:** For Polar artifacts, indicates which perspective the artifact reflects. *Neutral* for Consensus tiers; reference cluster (C_1) or opposing clusters for Polar tiers.
- **Risk if Acted Upon:** Epistemic risk of treating this artifact as authoritative without additional context.
 - *Low*: Positive Consensus—cross-cluster endorsement provides strong validation.
 - *Moderate*: Polar—valid perspective but reflects specific cluster’s view.
 - *High*: Negative Consensus—cross-cluster rejection signals significant concerns.

The orchestrator receives a list of these objects, enabling context-appropriate decisions. Crucially, even Negative Consensus artifacts carry information: “all perspectives rejected this” may be exactly what the user needs to understand.

8 Orchestrator Guidance

The orchestrator interprets Resonance States based on context and user needs. The following patterns illustrate common responses:

| Tier | Possible Orchestrator Responses |
|--------------------|--|
| Positive Consensus | Present with confidence. Cross-cluster endorsement provides strong validation. |
| Negative Consensus | Surface the rejection: “Both perspectives identified concerns with...” May prompt clarification from user or acknowledge limitation. |
| Positive Polar | Present with attribution: “From the [reference cluster] perspective...” Consider surfacing the disagreement. |
| Negative Polar | Present with attribution to opposing cluster. May indicate user’s framing aligns with a specific perspective. |
| Mixed results | When no Consensus exists, orchestrator may: request clarification, present multiple perspectives explicitly, or acknowledge genuine ambiguity. |

The key principle: **contestation is information, not failure**. When the protocol reveals disagreement, the orchestrator gains a principled basis for structured uncertainty—acknowledging what is known, what is contested, and what requires further input.

9 Algorithm

Algorithm 1 Resonant Consensus Protocol

Require: Agents \mathcal{A} partitioned into clusters $\mathcal{C} = \{C_1, \dots, C_n\}$ with reference cluster C_1

Require: Query q , thresholds θ (cluster approval), τ (consensus)

Ensure: List of Resonance State objects for orchestrator

```

1:
2: // Phase 1: Generate
3: for each  $a \in \mathcal{A}$  do
4:    $\omega_a \leftarrow a.\text{respond}(q)$ 
5: end for
6:  $\Omega \leftarrow \{\omega_a : a \in \mathcal{A}\}$ 
7:
8: // Phase 2: Vote
9: for each  $\omega \in \Omega$  do
10:   for each  $a \in \mathcal{A}$  where  $a \neq \text{author}(\omega)$  do
11:      $v(\omega, a) \leftarrow a.\text{approve}(\omega)$                                 ▷ Returns 0 or 1
12:   end for
13: end for
14:
15: // Phase 3: Classify
16: for each  $\omega \in \Omega$  do
17:   for each  $C_i \in \mathcal{C}$  do
18:      $r_i \leftarrow \frac{1}{|C_i|} \sum_{a \in C_i} v(\omega, a)$ 
19:      $\text{Approve}_i(\omega) \leftarrow \mathbf{1}[r_i \geq \theta_i]$ 
20:   end for
21:    $S(\omega) \leftarrow \{C_i : \text{Approve}_i(\omega) = 1\}$ 
22:    $\rho(\omega) \leftarrow |S(\omega)|/n$ 
23:   refApproves  $\leftarrow C_1 \in S(\omega)$ 
24:
25: if  $\rho(\omega) \geq \tau$  then
26:   tier  $\leftarrow \text{POSITIVECONSENSUS}$                                 ▷ All high- $\rho$  are positive
27: else if  $\rho(\omega) \leq 1 - \tau$  then
28:   tier  $\leftarrow \text{NEGATIVECONSENSUS}$ 
29: else if refApproves then
30:   tier  $\leftarrow \text{POSITIVEPOLAR}$ 
31: else
32:   tier  $\leftarrow \text{NEGATIVEPOLAR}$ 
33: end if
34:
35: Compute Score( $\omega$ ), assessment dimensions
36: Build Resonance State object
37: end for
38:
39: // Phase 4: Return
40: Group by tier, sort by Score descending within each tier
41: return list of Resonance State objects

```

10 Implementation Notes

10.1 Eliciting Votes

Prompt each agent:

Given your perspective, does this response represent sound reasoning you would endorse? Answer YES or NO.

Response to evaluate: [artifact text]

Map YES → 1, NO → 0.

10.2 Minimum Panel Size

Each cluster requires ≥ 2 agents for meaningful approval rates. With exactly 2 agents per cluster, $\theta = 0.5$ requires both to agree (unanimous within cluster).

Recommended minimum: 3 agents per cluster.

10.3 Parameter Defaults

| Parameter | Default | Meaning |
|-----------|---------|--|
| θ | 0.5 | Simple majority within cluster |
| τ | 0.6 | 60% of clusters required for Consensus |

11 Extensions

11.1 Per-Cluster Thresholds

Different clusters may warrant different approval thresholds. For example, a “safety reviewer” cluster might require $\theta_{\text{safety}} = 0.9$ (near-unanimous) while others use $\theta = 0.5$.

Set θ_i individually, or use the global default:

$$\theta_i = \theta_{\text{global}} \quad \text{for all } i \quad (\text{default behavior}) \quad (9)$$

11.2 Weighted Clusters

If some clusters are more authoritative, weight their approval:

$$\rho_{\text{weighted}}(\omega) = \frac{\sum_{C_i \in S(\omega)} w_i}{\sum_{i=1}^n w_i} \quad (10)$$

Where w_i is the weight of cluster C_i . This allows, e.g., domain experts to have more influence than general reviewers.

11.3 Iterative Refinement

For complex queries, iterate:

1. Run the protocol
2. Use Consensus artifacts as context for a second generation round
3. Re-evaluate and re-classify

This allows agents to refine positions based on verified common ground.

12 Summary

The Resonant Consensus Protocol classifies multi-agent outputs by cross-cluster resonance:

1. Partition agents into n adversarial clusters with designated reference cluster C_1
2. Collect binary approval votes on each artifact
3. Compute approval set $S(\omega)$ and resonance ratio $\rho(\omega)$
4. Classify into four tiers based on τ and reference cluster membership
5. Return Cross-Cluster Resonance State with multi-dimensional assessment

The key insight: **who endorses an artifact matters as much as how many endorse it.** An artifact with 60% approval is Positive Consensus if that support spans clusters, but Polar if concentrated in one. This distinction enables orchestrators to separate robust insights from factional positions—and to represent genuine uncertainty rather than gambling on best guesses.

Quick Reference

| Symbol | Definition |
|-------------------------------------|---|
| $\mathcal{C} = \{C_1, \dots, C_n\}$ | Set of adversarial clusters (C_1 = reference) |
| $v(\omega, a)$ | Binary vote: agent a approves artifact ω |
| θ_i | Approval threshold for cluster C_i |
| $S(\omega)$ | Approval set: clusters that approve ω |
| $\rho(\omega) = S(\omega) /n$ | Resonance ratio |
| τ | Consensus threshold |

| Tier | Condition | Contestation | Risk |
|--------------------|--|--------------|----------|
| Positive Consensus | $\rho \geq \tau$ | Low | Low |
| Negative Consensus | $\rho \leq 1 - \tau$ | Low | High |
| Positive Polar | $1 - \tau < \rho < \tau, C_1 \in S$ | High | Moderate |
| Negative Polar | $1 - \tau < \rho < \tau, C_1 \notin S$ | High | Moderate |

References

References

- [1] Kalai, A.T., Ramchandran, S.K., Vempala, S.S., & Zhang, R.J. (2025). Why language models hallucinate. *OpenAI Research*. <https://openai.com/index/why-language-models-hallucinate/>
- [2] Wu, J., et al. (2025). Mitigating LLM hallucination via behaviorally calibrated reinforcement learning. *arXiv preprint arXiv:2512.19920*.
- [3] Farquhar, S., Kossen, J., Kuhn, L., & Gal, Y. (2024). Detecting hallucinations in large language models using semantic entropy. *Nature*, 630, 625–630.
- [4] Huang, Q., et al. (2025). Medical hallucination in foundation models and their impact on healthcare. *medRxiv preprint*.

- [5] Sharma, M., et al. (2024). Sycophancy in large language models: Causes and mitigations. *arXiv preprint arXiv:2411.15287*.
- [6] Fanous, A., Goldberg, J., et al. (2025). SycEval: Evaluating LLM sycophancy. *arXiv preprint arXiv:2502.08177*.