The Resonant Consensus Protocol

Cross-Cluster Agreement Classification for Multi-Agent Systems

Version 4.0

1 Overview

When multiple LLM agents with different perspectives respond to a query, how should an orchestrator decide which responses represent genuine consensus versus contested positions?

The Resonant Consensus Protocol classifies artifacts based on **cross-cluster agreement**: which adversarial groups approve, and how many. This produces a **Contextual Superposition**—a structured summary that tells the orchestrator not just what was said, but how it resonated across perspectives.

2 General Framework

2.1 Components

System Elements

- **Agents** $\mathcal{A} = \{a_1, \ldots, a_m\}$ — LLM instances with distinct system prompts
- **Clusters** $\mathcal{C} = \{C_1, \ldots, C_n\}$ — Partition of agents into n adversarial groups
- **Artifacts** $\Omega = \{\omega_1, \ldots, \omega_k\}$ — Responses generated by agents
- **Votes** $v(\omega, a) \in \{0, 1\}$ — Agent a's approval of artifact ω

Clusters are defined at design time based on system prompt orientation. Each cluster represents a distinct perspective designed to stress-test artifacts from other viewpoints (e.g., advocate vs. critic, technical vs. business, optimist vs. skeptic).

2.2 Cluster Approval

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

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

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).

2.3 The Approval Set

The **approval set** is the set of clusters that approve an artifact:

$$S(\omega) = \{C_i \in \mathcal{C} : \text{Approve}_i(\omega) = 1\} \tag{2}$$

This captures *which* clusters approved, not just how many.

2.4 Agreement Ratio

The **agreement ratio** normalizes the approval set cardinality:

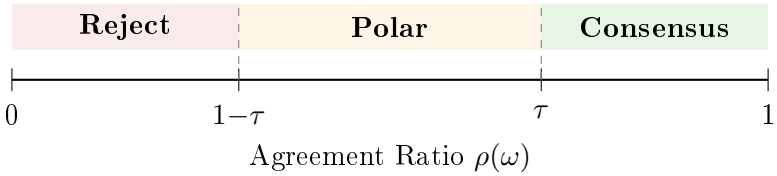
$$\rho(\omega) = \frac{|S(\omega)|}{n} \in [0, 1] \tag{3}$$

This represents the fraction of clusters that approve the artifact.

3 Classification

Artifacts are classified into three tiers based on the agreement ratio $\rho(\omega)$ and a threshold parameter $\tau \in (0.5, 1]$:

$$\text{Tier}(\omega) = \begin{cases} \textbf{Consensus} & \text{if } \rho(\omega) \geq \tau \\ \textbf{Polar} & \text{if } 1 - \tau < \rho(\omega) < \tau \\ \textbf{Reject} & \text{if } \rho(\omega) \leq 1 - \tau \end{cases} \tag{4}$$



The threshold τ controls the boundaries symmetrically:

- Higher τ (e.g., 0.8) → stricter consensus requirement, wider Polar band
- Lower τ (e.g., 0.6) → more permissive consensus, narrower Polar band
- $\tau = 0.5$ → no Polar tier (binary: Consensus or Reject)
- $\tau = 1.0$ → only unanimous approval counts as Consensus

3.1 Tier Definitions

Consensus — $\rho(\omega) \geq \tau$

A strong majority of clusters approve. This artifact resonates across adversarial perspectives and represents robust agreement.

Action: GROUND — Use as primary context for response synthesis.

Special case: When $|S(\omega)| = n$ (all clusters approve), this is **Full Consensus**—the highest confidence level. The orchestrator may flag this distinction.

Polar — $1 - \tau < \rho(\omega) < \tau$

Mixed approval—some clusters approve, others reject. This artifact is contested; it may represent a valid perspective but not cross-cutting agreement.

Action: CONTEXTUALIZE — Include with attribution to approving clusters.

Reject — $\rho(\omega) \leq 1 - \tau$

A strong majority of clusters reject. This artifact fails cross-examination and should be discarded.

Action: EXCLUDE — Do not include in response synthesis.

3.2 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, not an error—it simply means 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:

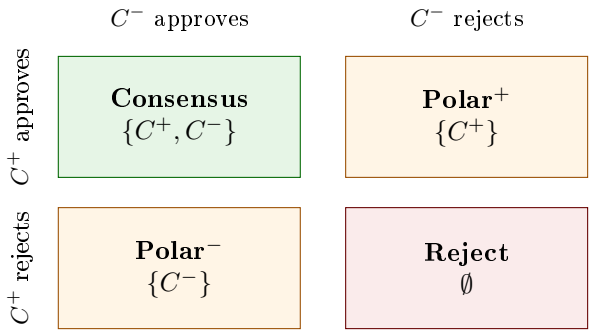
$S(\omega)$	$\lvert S\rvert$	ρ	Interpretation
$\{C^+, C^-\}$	2	1.0	Both clusters approve
$\{C^+\}$	1	0.5	Only advocates approve
$\{C^-\}$	1	0.5	Only critics approve
$\emptyset$	0	0.0	Neither approves

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**:

$$\text{Score}(\omega) = \frac{1}{|\mathcal{A}| - 1} \sum_{a \neq \text{author}(\omega)} v(\omega, a) \tag{5}$$

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 agreement:

$$\text{Score}_{\text{balanced}}(\omega) = \text{Score}(\omega) \times (1 - \sigma_S(\omega)) \tag{6}$$

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

$$\sigma_S(\omega) = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (r_i(\omega) - \bar{r}(\omega))^2} \tag{7}$$

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 $|S(\omega) \backslash \{C_i\}|$ — 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 Contextual Superposition

The protocol outputs a **Contextual Superposition** for each artifact—a structured summary for the orchestrator:

Superposition Object

```
Superposition() = {
    artifact: ,
    approval_set: S(),
    agreement_ratio: (),
    tier: Consensus | Polar | Reject,
    score: Score(),
    author_cluster: C_i,
    is_persuasive: boolean,
    full_consensus: boolean  // true if |S| = n
}
```

The orchestrator receives a list of these objects, grouped by tier and sorted by score.

8 Orchestrator Actions

Tier	Action	Guidance
Consensus (Full)	GROUND	Highest confidence. Lead with these points.
Consensus (Partial)	GROUND	High confidence. Note dissenting clusters if relevant.
Polar	CONTEXTUALIZE	Include with attribution: "From [cluster] perspective..."
Reject	EXCLUDE	Do not reference.
Persuasive	EMPHASIZE	Strong bridge-building points.

9 Algorithm

Algorithm 1 Resonant Consensus Protocol

Require: Agents $\mathcal{A}$ partitioned into clusters $\mathcal{C} = \{C_1, \ldots, C_n\}$
Require: Query q, thresholds θ (cluster approval), τ (consensus)
Ensure: List of Superposition objects for orchestrator

// **Phase 1: Generate**
for each $a \in \mathcal{A}$ **do**
    $\omega_a \leftarrow a.\text{respond}(q)$
end for
$\Omega \leftarrow \{\omega_a : a \in \mathcal{A}\}$

// **Phase 2: Vote**
for each $\omega \in \Omega$ **do**
    for each $a \in \mathcal{A}$ where $a \neq \text{author}(\omega)$ **do**
        $v(\omega, a) \leftarrow a.\text{approve}(\omega)$ ▷ Returns 0 or 1
    end for
end for

// **Phase 3: Classify**
for each $\omega \in \Omega$ **do**
    for each $C_i \in \mathcal{C}$ **do**
        $r_i \leftarrow \frac{1}{|C_i|} \sum_{a \in C_i} v(\omega, a)$
        $\text{Approve}_i(\omega) \leftarrow \mathbf{1}[r_i \geq \theta_i]$
    end for
    $S(\omega) \leftarrow \{C_i : \text{Approve}_i(\omega) = 1\}$
    $\rho(\omega) \leftarrow |S(\omega)|/n$

    if $\rho(\omega) \geq \tau$ **then**
        tier $\leftarrow$ Consensus
    else if $\rho(\omega) > 1 - \tau$ **then**
        tier $\leftarrow$ Polar
    else
        tier $\leftarrow$ Reject
    end if

    Compute Score(ω)
    Build Superposition object
end for

// **Phase 4: Return**
Sort by tier (Consensus > Polar > Reject), then by Score descending
return list of Superposition 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 $\to 1$, NO $\to 0$.

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:

$$\theta_i = \theta_{\text{global}} \quad \text{for all } i \quad \text{(default behavior)} \tag{8}$$

11.2 Weighted Clusters

If some clusters are more authoritative, weight their approval:

$$\rho_{\text{weighted}}(\omega) = \frac{\sum_{C_i \in S(\omega)} w_i}{\sum_{i=1}^{n} w_i} \tag{9}$$

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 agreement:

1. Partition agents into n adversarial clusters
2. Collect binary approval votes on each artifact
3. Compute approval set $S(\omega)$ and agreement ratio $\rho(\omega)$
4. Classify into Consensus / Polar / Reject based on threshold τ
5. Return Contextual Superposition to orchestrator

The key insight: **who agrees matters as much as how many agree**. An artifact with 60% approval is Consensus if that support spans clusters, but Polar if concentrated in one. This distinction enables orchestrators to separate robust insights from tribal bias.

Quick Reference

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

Condition	Tier	Action
$\rho \geq \tau$	Consensus	GROUND
$1 - \tau < \rho < \tau$	Polar	CONTEXTUALIZE
$\rho \leq 1 - \tau$	Reject	EXCLUDE