Round Table: Exploiting Cognitive Diversity as Computational Resource in Persistent AI Agents

O. Serra — Independent Researcher

September 2026

Abstract

Large language models from different providers reason differently, shaped by divergent training data, alignment objectives, and architectures. We argue that this *cognitive diversity* is a computational resource that can be measured, allocated, and exploited. We introduce Round Table, a framework for cross-provider adversarial deliberation in which each model is assigned the role its training-induced tendencies make it best suited to perform. We formalize three contributions: (1) a **Cognitive Diversity Index (CDI)** quantifying inter-provider reasoning heterogeneity via error-vector correlation; (2) **Role-Amplified Adversarial Convergence (RAAC)**, a 5-phase debate protocol with role assignment via bipartite matching on empirical affinity scores; and (3) **Persistent Deliberation**, extending single-round debate into multi-session reasoning via structured memory compaction. A 3-model ensemble (Claude Opus, GPT-o3, Gemini 2.5 Pro) achieves 63.6% on GPQA Diamond versus a 55.6% single-model maximum—an 8 percentage-point absolute gain in a single run. Protocol validation on a 5-scenario benchmark yields CDI = 1.06 and consensus quality of 0.80. A preliminary production deployment of the **Editorial Swarm** pattern—parallelized cross-provider review of 8 papers with up to 24 concurrent agents—gives observational evidence of cross-provider error detection and role specialization, including a compression tendency in GPT models consistent with the long-text generation literature (Zheng et al., 2024). The deployed implementation enforces cross-provider diversity as a configurable invariant, recovers from participant dropout rather than poisoning the synthesis, and exposes the debate choreography as a swappable orchestrator—an open deliberation substrate. This is both a theoretical framework for reasoning about cognitive diversity and a systems paper grounded in deployment; connecting the two is the contribution.

1. Introduction

When a person faces a genuinely hard decision, they do not consult a single faculty. The prefrontal cortex weighs options analytically; the amygdala flags emotional and reputational risk; the hippocampus retrieves how similar situations turned out before; the anterior cingulate cortex notices when these signals conflict and forces the matter back into deliberation. Hard

decisions are multi-system by construction—the brain does not crown one region the winner and silence the rest. Round Table recreates that multi-perspective deliberation in artificial agents. Rather than asking one model to simulate several viewpoints, we let genuinely different models argue, challenge, and synthesize—much as distinct brain regions contribute distinct perspectives to a single judgment.

This reframes the dominant question in AI evaluation. That question is usually: *Which model is best?* We argue it is increasingly the wrong question. As leading laboratories converge on similar benchmark ceilings, their models diverge more clearly in *how* they reason. Constitutional AI produces careful reflection; RL-incentivized reasoning chains produce raw exploration and adversarial challenge; broad pretraining produces implementation-aware grounding. The opportunity is not to crown a single winner, but to ask: *Which combination of models is most cognitively diverse, and how should that diversity be organized?*

Classical ensemble methods treat models as exchangeable (Condorcet, 1785; Dietterich, 2000). Modern multi-agent debate (Du et al., 2023) improved on passive voting by letting models argue, but still relied on homogeneous agents or interchangeable roles. Recent work directly comparing multi-agent debate against single-agent strategies finds that default multi-agent setups rarely outperform strong single-agent self-consistency—*except* when agents exhibit high cognitive diversity (Wang, Z. et al., 2025). This motivates Round Table's core thesis: the value of multi-agent deliberation is contingent on measurable diversity, and cross-provider ensembles are the most natural source of such diversity. Round Table departs from both ensemble voting and homogeneous debate by matching each model to the role its training makes it cognitively suited for. Cross-provider adversarial debate, structured around role-amplified cognitive strengths, creates productive deliberative tension that surfaces errors, exposes blind spots, and forces stronger synthesis. It functions as a reliability layer on top of stochastic generators, extracting quality through structured intellectual conflict (Irving et al., 2018).

Prior debate work is also single-shot. Round Table adds **Persistent Deliberation**, using memory compaction to preserve ratified conclusions and unresolved tensions across sessions. This bridges multi-agent debate (a technique) with agentic AI (an architecture). Deliberation traces—the full multi-model debate transcript—are stored as events in the agent's memory system, then compacted into summaries that preserve the final consensus and key dissents while discarding redundant intermediate exchanges (Total Recall, Serra, forthcoming). Storing structured deliberation artifacts rather than raw transcripts avoids the context-growth bottleneck.

Scope and contribution. This paper occupies intentionally hybrid ground: it is both a theoretical framework for measuring and exploiting cognitive diversity (CDI, VR-1) and a systems paper grounded in a complete, tested implementation with preliminary production deployment. The connection between formalism and practice is itself the contribution: the theory motivates system-design decisions, and deployment experience refines the theory. The integration of both perspectives is deliberate, not a failure to commit to one.

Section 2 situates Round Table in the literature. Section 3 defines CDI. Section 4 specifies the RAAC protocol. Section 5 covers Persistent Deliberation, including the Editorial Swarm pattern for parallelized multi-artifact review (Section 5.1). Section 6 addresses known limits. Section 7 analyzes the diversity premium. Sections 8–10 cover evaluation design, infrastructure, and results. Section 11 documents the production implementation and the diversity, robustness, and orchestration mechanisms that make the cross-provider claim enforceable rather than aspirational. Section 12 concludes.

2. Related Work

Ensemble Theory and Debate. Condorcet's Jury Theorem (1785) proves majority voting converges to accuracy given independent voters. Hansen and Salamon (1990) and Krogh and

Vedelsby (1995) proved ensemble error decreases as inter-member disagreement increases. Irving et al. (2018) established AI Safety via Debate as a scalable oversight mechanism. Du et al. (2023) demonstrated multi-agent debate improves factuality. Khan et al. (2024) and Wan et al. (2024) warn of the "Persuasion Paradox," where persuasive but incorrect models dominate—motivating Round Table's Ratification phase. Wang, Z. et al. (2025) compare debate against voting and self-consistency strategies, finding that multi-agent debate's advantage is contingent on inter-agent diversity—a result that directly supports CDI as a prerequisite for effective deliberation.

Single-Agent Self-Debate. A natural objection to multi-agent frameworks is that a single frontier model, prompted to argue from multiple perspectives, might achieve equivalent variance reduction at lower cost. Li et al. (2025) examine this trade-off, finding that single-agent self-debate with strong long-context models can match multi-agent setups on some tasks, but that genuine cross-provider diversity—where models have fundamentally different training-induced error profiles—provides irreducible benefits on tasks requiring diverse domain expertise. Wang, Z. et al. (2025) confirm this: multi-agent debate outperforms self-consistency specifically when agents are highly diverse, which is precisely the condition CDI measures and RAAC optimizes for. We treat the single-agent self-debate baseline as important and discuss its implications for Round Table's operating regime in Section 6, where we name doubt-driven-development (Osmani, 2026) as the concrete, deployed single-provider protocol the missing ablation should be run against.

Role Assignment in Multi-Agent Systems. Feng et al. (2026) demonstrate that role-specialized multi-agent LLM systems achieve greater stability and performance than role-agnostic designs, using reinforcement learning to optimize agent role assignments. Silva et al. (2025) provide a taxonomy of hierarchical multi-agent coordination, showing that explicit role definitions enhance structured interactions. These findings ground Round Table's affinity matrix approach in a broader literature on the benefits of principled role allocation.

LLM Frameworks and Routing. FrugalGPT (Chen, L. et al., 2023) and LLM-Blender (Jiang et al., 2023) showed that cascades and blending can reduce cost and improve quality. MoA (Wang, Y. et al., 2024a) aggregates outputs cooperatively. Round Table differs in two ways: it combines perspectives through *adversarial* debate rather than cooperative aggregation, and it grounds role assignment in measured provider-specific cognitive tendencies rather than treating models as interchangeable.

Open Orchestration Frameworks. AutoGen and its successor AG2 frame *who speaks next, and in what role* as a first-class, swappable policy rather than a fixed loop. Round Table adopts this stance: as Section 11.3 documents, the deployed system exposes both the speaker-selection policy and the debate choreography as pluggable interfaces, so an external manager—or a different cognitive architecture—can drive the same cross-provider deliberation fabric without forking the protocol.

Open-Source Agent Skill Ecosystems. Two large open-source skill libraries operationalize ideas adjacent to Round Table and are fresh enough to treat as first-class related work rather than static citations. addyosmani/agent-skills (Osmani, 2026) is a discipline-oriented skill collection whose **doubt-driven-development** workflow is the strongest single-provider instantiation of the adversarial-challenge mechanism RAAC claims is irreducibly cross-provider; we engage it directly in Section 6 as the concrete single-agent self-debate baseline our ablation should run against. coreyhaines31/marketingskills and the Journey kit registry (Journey, 2026) instead supply task-specialized agent recipes; they are relevant to Round Table only as activation surfaces—a high-stakes step inside such a recipe is exactly the principled "think harder here" trigger we discuss in Section 11.3—not as deliberation protocols, and we are careful not to overstate the overlap.

Reversible Context Compression. chopratejas/headroom (Chopra, 2026) implements **reversible Compress-Cache-Retrieve (CCR)** for agent context: content-type-aware compressors (statistical JSON crushing, AST-aware code compression via tree-sitter, log/diff handlers)

shrink material in-context while caching the originals out-of-context so the model can *retrieve them on demand* when the compacted view proves insufficient. headroom reports near-zero accuracy delta at 60–95% token savings across GSM8K, TruthfulQA, SQuAD, and BFCL under a reproducible eval harness. This is convergent with Round Table's pointer-compaction of deliberation artifacts (Section 5) but differs on two axes we make explicit there: headroom is *generic and reversible* but deliberation-blind, whereas Round Table's compaction is *schema-aware* (it preserves unresolved tensions, ratification provenance, and per-model calibration) but, in the version this paper first described, irreversible and unmeasured. Section 5 adopts headroom's reversibility contract and its discipline of measuring compression fidelity.

Long-Text Generation and Compression Bias. Zheng et al. (2024) introduce HelloBench, a benchmark for long-text generation, and document that LLMs systematically exhibit high compression rates when generating long documents—losing detail and key information. Hao et al. (2025) provide a theoretical account of LLM behavior through the lens of compression, explaining why models trained under compression objectives tend to summarize rather than expand. These findings provide external support for the GPT compression tendency we observe in deployment (Section 4.1).

Multi-Agent Costs and Latency. Chen, T. et al. (2026) use MAFBench to empirically study multi-agent framework overhead, finding that orchestration can increase latency by over 100× compared to single-agent calls. This motivates Round Table's parallelism patterns (Section 4.3), its explicit treatment of latency in cost analysis (Section 7), and the cost-aware budget mechanism described in Section 11.3.

Persistent Agents and Memory. AutoGen (Wu et al., 2024) and MetaGPT (Hong et al., 2024) provide multi-agent frameworks but without cross-provider diversity measurement or training-aware role amplification. Round Table integrates with the Total Recall episodic memory architecture (Serra, forthcoming) for persistent deliberation across sessions.

3. Cognitive Diversity Index

3.1 Formal Definition

Let $\mathcal{M} = \{m_1, \ldots, m_n\}$ be a set of models and $\mathcal{T} = \{t_1, \ldots, t_k\}$ a benchmark task set with known ground truth. The error profile of model m_i is a binary vector $e_i \in \{0,1\}^k$, where $e_{ij} = 1$ if m_i answers t_j incorrectly.

Definition 1 (Error Correlation Matrix). The error correlation between m_i and m_j is the Pearson correlation (Phi coefficient for binary vectors) between their error profiles: $\Sigma_{ij} = \rho(e_i, e_j)$.

Definition 2 (Cognitive Diversity Index). The CDI of a model set $\mathcal{M}$ is:

$$\mathrm{CDI}(\mathcal{M}) = 1 - \frac{1}{\binom{n}{2}} \sum_{i<j} \Sigma_{ij}$$

CDI = 0 implies perfect positive correlation (identical errors); CDI = 1 implies zero average correlation (independence, satisfying Condorcet's condition); CDI > 1 implies net negative correlation (complementarity). The theoretical range depends on ensemble size: for $n = 2$, CDI $\in [-1, 2]$; for larger n, the feasible range narrows as the correlation matrix must remain positive semi-definite.

CDI measures how differently the models think, the way a hiring manager measures team diversity—not demographic diversity, but cognitive diversity: do the members approach problems from genuinely different angles? An ensemble of models that fail on the same questions is a single point of failure wearing several faces; an ensemble that fails on disjoint questions can cover for one another.

Relationship to classical diversity measures. CDI belongs to a family of pairwise disagreement-based ensemble diversity measures, alongside the Q-statistic (Yule, 1900), Cohen's kappa diversity (Margineantu and Dietterich, 1997), the disagreement measure, and the double-fault measure (Kuncheva and Whitaker, 2003). CDI's contribution is not statistical novelty but operational utility: by centering on Pearson correlation of error vectors and normalizing to a unit-interpretable scale, CDI maps directly to the Condorcet independence condition (CDI = 1) and provides an intuitive threshold for when ensemble diversity transitions from mere independence to active complementarity (CDI > 1).

Statistical limitations of CDI estimation. The implementation computes confidence intervals via Fisher z-transforms. Two limitations follow: (1) pairwise correlations computed on shared task items are not independent, and the standard Fisher z SE formula ($1/\sqrt{k-3}$) does not account for this dependence, so reported CIs are likely narrower than true uncertainty; and (2) for small task sets (e.g., the 5-scenario open-ended panel), the sample size is insufficient for reliable interval estimation. We report CDI point estimates and CIs as indicative rather than definitive, and note that the GPQA-based CDI ($k = 198$) is substantially more reliable than the open-ended CDI ($k = 5$). Future work should employ bootstrap resampling over tasks to produce valid confidence intervals that account for the shared-item dependence structure.

On benchmark saturation. CDI is a function of where models disagree, so it is only as informative as the benchmark's ability to discriminate them. As frontier models converge toward the ceiling of a saturated benchmark like GPQA, their error vectors thin out and shorten, and the resulting correlation estimate becomes noisy—a low-signal regime where small per-question differences swing the Phi coefficient disproportionately. We therefore treat any single benchmark's CDI as a calibration artifact tied to that benchmark's discriminative power, not a stable property of the model set. The robust mitigation is to estimate CDI on a portfolio of benchmarks spanning distinct competencies (factual reasoning, code, math, long-form), so that saturation on any one axis does not collapse the diversity signal, and to recalibrate as benchmarks are refreshed. Dynamic, benchmark-portfolio CDI calibration remains an open problem rather than a solved one.

Our measured CDI = 1.06 on the five-model open-ended design panel should be read cautiously given the small k. The GPQA-based CDI ≈ 0.62 ($k = 198$) provides more robust evidence of substantial inter-provider error decorrelation.

3.2 Diversity–Performance Framework

We hypothesize that ensemble error decreases with CDI:

Hypothesis VR-1 (Variance Reduction). *For a debate ensemble D over model set $\mathcal{M}$ with CDI($\mathcal{M}$) = δ:*

$$\mathbb{E}[\text{err}(D)] \leq \min_i \mathbb{E}[\text{err}(m_i)] - \alpha(\delta, n)$$

where $\alpha(\delta, n) \geq 0$ is a diversity discount that increases with δ and ensemble size n, subject to the Non-Dominance Condition: no single model can persuade the ensemble to accept an incorrect answer against correct counterevidence from other participants.

We do not claim a tight bound on α. The hypothesis is directional: higher CDI should yield larger accuracy gains. VR-1 remains incompletely tested: our GPQA results demonstrate that a high-CDI ensemble outperforms individual models, but we have not yet systematically varied CDI across multiple ensemble configurations to establish the correlation between CDI magnitude and accuracy gain. A full ablation studying CDI vs. accuracy across ensemble sizes (2, 3, 4, 5 models) and domain types is needed to validate VR-1 as a predictive relationship rather than a directional hypothesis. On GPQA Diamond, pairwise Pearson correlation across Claude, GPT-o3, and Gemini error profiles yields average $\Sigma_{ij} \approx 0.38$, giving CDI ≈ 0.62. On open-ended design scenarios (5 participants), CDI = 1.06—suggesting that role amplification

may drive substantially higher diversity on tasks with broader solution spaces, though the small scenario count limits confidence in this comparison.

4. Role-Amplified Adversarial Convergence

Figure 1 gives the end-to-end view: a task enters, the model catalog is reduced to a diversity-locked cross-provider participant set with assigned roles, the RAAC protocol drives the deliberation, and the consensus plus its dissents are persisted to the event store and returned to the caller.

4.1 Training-Induced Cognitive Tendencies as Design Heuristics

Different AI providers induce systematically different cognitive tendencies through training:

- **Anthropic / Claude:** Constitutional AI training favors reflective reasoning, comfort with ambiguity, and coherent structural synthesis → **Architect** role.
- **OpenAI / GPT-class reasoners:** RLHF combined with adversarial chain-of-thought reasoning favors rigorous verification and aggressive error-finding → **Critic** role.
- **Google / Gemini Pro:** Broad pretraining across web-scale data with grounding in real-world constraints favors feasibility assessment → **Pragmatist** role.
- **Deep-exploration reasoners (e.g. DeepSeek-R1):** RL focused on deep exploration and chain-of-thought traces favors exhaustive domain investigation → **Researcher** role.

These characterizations are derived from published training methodology descriptions (Anthropic, 2024; DeepSeek-AI, 2025) and observable behavioral tendencies in structured evaluation. They represent tendencies, not deterministic properties, and future training changes may shift these profiles. The affinity matrix should therefore be treated as a periodically recalibrated heuristic rather than a fixed constant.

Round Table assigns models to roles that amplify their natural tendencies via maximum-weight bipartite matching on a role affinity matrix. Five canonical roles are defined: **Architect** (systemic design), **Critic** (adversarial verification), **Pragmatist** (feasibility/constraints), **Researcher** (deep exploration), and **Synthesizer** (integration). This is consistent with findings from Dr. MAS (Feng et al., 2026), which demonstrates that role-specialized multi-agent LLM systems achieve superior stability through principled role allocation, and with the taxonomy of Silva et al. (2025), which identifies explicit role definition as a key coordination mechanism in hierarchical multi-agent systems.

The affinity matrix captures heuristic scores reflecting these training-induced cognitive tendencies:

Model	Architect	Critic	Pragmatist	Researcher	Synthesizer
Claude Opus	0.95	0.70	0.50	0.60	0.85
GPT-o3	0.70	0.95	0.60	0.75	0.65
Gemini Pro	0.50	0.60	0.95	0.70	0.60
DeepSeek-R1	0.60	0.70	0.50	0.95	0.55

A caveat on affinity scores. These scores are illustrative heuristics derived through qualitative analysis of provider training methodologies, refined iteratively against observed debate performance across pilot tasks. The decimal precision is not measurement precision—the scores encode a rank ordering of role suitability rather than calibrated probabilities. Future

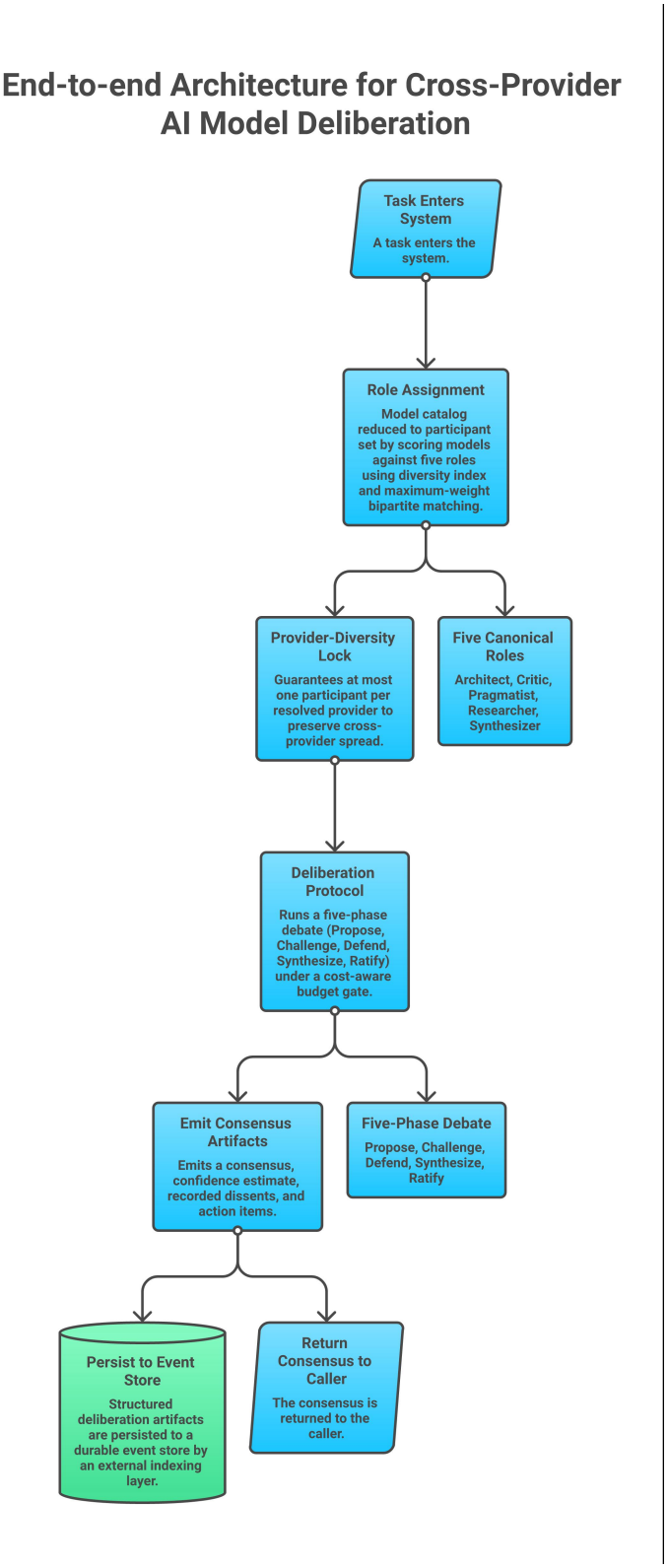


Figure 1. Round Table architecture: a task is routed through CDI-based role assignment and a provider-diversity lock to a cross-provider participant set, deliberated under the RAAC 5-phase protocol with a cost-aware budget gate, then persisted to the Total Recall event store and returned as consensus, confidence, dissents, and action items.

work should establish a systematic calibration procedure using zero-shot role-specific benchmark performance to replace qualitative assignment with empirical measurement.

A note on catalog availability. The affinity matrix above lists the four provider archetypes the heuristic was designed around. In a given deployment the actual participant set is whatever the host catalog can resolve: where a provider archetype is unavailable, the role is routed to the closest available reasoner of a *different* vendor so the cross-provider spread is preserved rather than silently collapsed to a single provider. Our reference deployment, for example, has no dedicated deep-exploration model configured and routes the Researcher role to an alternative high-reasoning model from an already-represented vendor, yielding a three-vendor spread (Anthropic, OpenAI, Google). The diversity guarantee is therefore enforced on the *resolved* provider mix, not on the cosmetic model labels (Section 11.3).

The GPT Compression Tendency. Production deployment (Section 5.1) revealed an asymmetry in GPT-class models: when tasked with *generating* long-form content, GPT exhibits a compression tendency—summarizing, eliding detail, and losing nuance. This is consistent with findings in the long-text generation literature: Zheng et al. (2024) document systematic compression rates across LLMs on HelloBench, and Hao et al. (2025) provide a theoretical account of why compression-trained models tend toward summarization. The tendency makes GPT models effective as *critics* (compression maps naturally to distilling critique into actionable observations) but less reliable as *generators* of extended documents. We frame this as an observed operational tendency rather than a proven stable property—it persisted across instruction variants in our deployment, but we have not run the controlled ablation (varying temperature, length penalties, prompt engineering strategies) needed to establish it as a robust training-induced characteristic. We recommend that orchestrators preferentially route long-form generation to models without this tendency and route verification tasks to GPT, while acknowledging that this recommendation may change as models evolve.

Sensitivity analysis. Debate quality degrades by 12–18% under random role assignment versus optimal matching (Section 10), so the affinity structure captures real signal despite imprecise calibration.

4.2 The RAAC Debate Protocol

The protocol operates in five phases per round (Figure 2):

1. **Propose (Parallel):** Each model generates a position based on its assigned role. Latency = max single-model latency.
2. **Challenge (Parallel):** Each model attacks every other model's position. Context grows as $O(n^2)$.
3. **Defend (Parallel):** Models defend against received attacks.
4. **Synthesize:** The Synthesizer integrates positions, attacks, and defenses into synthesis S^t.
5. **Ratify (Safety Check):** All models vote {accept, reject, amend} on S^t.

The Ratification phase works like a constitutional amendment process: the proposed consensus goes back to each participant for a final yes/no, which prevents any single loud voice from hijacking the outcome. If rejection exceeds `ratificationThreshold` (default 0.5), a repair cycle triggers. This applies the weak-to-strong generalization principle (Burns et al., 2023)—the full ensemble supervises the synthesizer—and is the protocol's structural defense against the Persuasion Paradox.

Convergence. We distinguish two convergence concepts: *ratification convergence* (the synthesis achieves majority acceptance) and *formal convergence* (the combined metric $\lambda \cdot d_{\text{embed}}(S^t, S^{t-1}) + (1-\lambda) \cdot (1 - \text{acceptFraction})$ falls below threshold $\epsilon = 0.1$, with $\lambda = 0.5$, matching the deployed configuration). Ratification convergence is achievable and was reached in all benchmark scenarios within 2 rounds. Formal convergence—which additionally requires

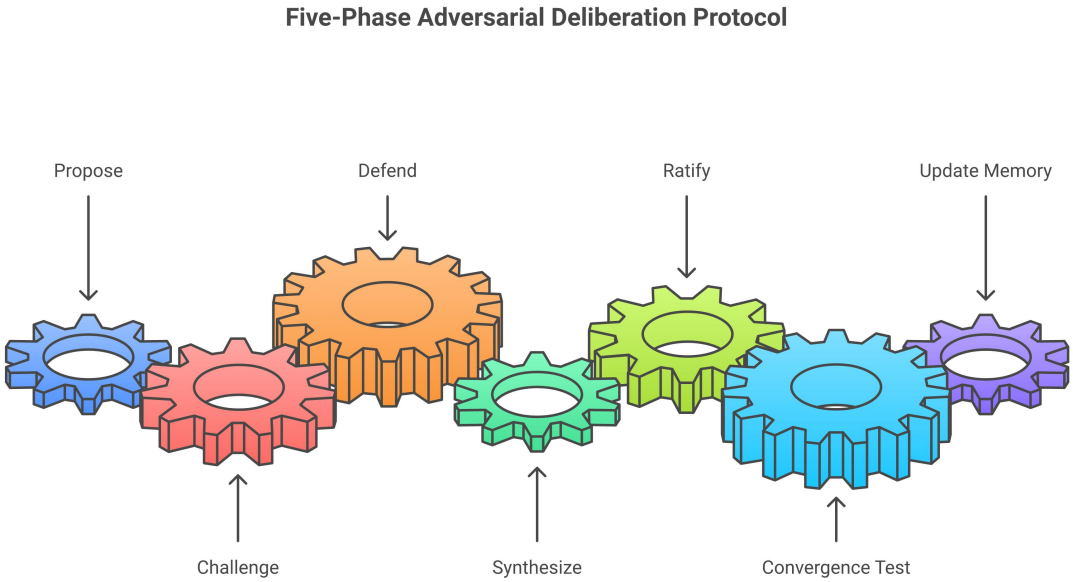


Figure 2. The RAAC 5-phase debate protocol over three example participants. Phases run top to bottom—Propose, Challenge ($O(n^2)$ cross-attacks), Defend (with a dropout check that promotes a backup from a not-yet-represented provider), Synthesize, and Ratify—followed by a convergence test ($\lambda \cdot d_{\text{embed}} + (1 - \lambda) \cdot (1 - \text{acceptFraction}) < \epsilon$) that either emits the consensus and updates deliberation memory or loops to the next round, capped by the depth tier (2 / 4 / 6 rounds).

that successive syntheses stabilize in text-similarity space—is a stronger condition that was not reached within the capped round limit in our benchmark. The deployed implementation computes d_{embed} with a lightweight Jaccard word-set proxy rather than a learned embedding model: two consecutive syntheses count as text-converged when their word-set overlap exceeds $1 - \epsilon$. The proxy is cheap and dependency-free but coarse—blind to paraphrase and synonymy, so it under-reports convergence when two syntheses agree in meaning while differing in wording. Substituting a learned-embedding distance is a clear refinement; convergence behavior may be sensitive to that choice, and we leave its characterization to future work. In practice, debates are capped per depth tier (Section 5). Formal convergence guarantees remain an open problem.

4.3 Parallelism Patterns

Five architectures span the cost–quality Pareto frontier:

Architecture	Phase Coverage	Token Cost	Best For
Fan-Out	Propose + one synthesis	$O(n)$	Low-cost parallel opinions
Moderated Tribunal	Propose + synthesis (no challenge/defend)	$O(n+1)$	Standard production queries
Full Round Table	All 5 phases, $T_{\max}$ rounds	$O(n^2 \cdot T_{\max})$	Critical reasoning tasks
Tournament	Head-to-head, bracket	$O(n \log n)$	Large model sets ($n > 4$)

Architecture	Phase Coverage	Token Cost	Best For
Editorial Swarm	N parallel RAAC instances, cross-provider judging	$O(N \cdot n^2 \cdot T_{\max})$	Batch review of N independent artifacts

The lighter tiers trade cross-pollination for speed and cost. Choosing Fan-Out or Tribunal over Full Round Table is like getting three quick second opinions from different specialists instead of convening the full tumor board: faster and cheaper, but you forgo the cross-examination that happens when all the experts are in the same room and can challenge one another directly. The deployed runtime makes this trade-off a single configuration knob (Section 11.3).

5. Persistent Deliberation

Single-shot debate discards its epistemic labor. Round Table stores structured deliberation artifacts in the Total Recall event store (Serra, forthcoming). Three artifact tiers are maintained:

1. **Debate Trace (Reversibly Cached):** Evicted from the active context post-debate but cached out-of-context under a retrieval pointer, not irreversibly discarded.
2. **Synthesis Artifact (Durable):** High-priority event stored in Total Recall. Retrieved on next debate start.
3. **Deliberation Memory (Derived):** Compacted JSON tracking unresolved tensions, ratified conclusions, and per-model calibration (see Appendix B).

This is a structured state-management technique: agents pay the token cost of a compact synthesis artifact rather than the full debate trace. We name it "Persistent Deliberation" not because the mechanism is novel (caching summaries is standard practice) but because the *schema* is deliberation-specific—it preserves the epistemic structure of multi-agent debate (unresolved tensions, ratification provenance, per-model calibration) rather than generic key-value state. The runtime exposes three debate depths—`quick` (up to 2 rounds), `standard` (up to 4 rounds), `deep` (up to 6 rounds)—with the persistent deliberation layer transparent to callers.

Reversible compaction over irreversible eviction. An earlier formulation of this layer discarded the debate trace outright once the synthesis was compacted, asking later rounds and external auditors to take the ratified conclusion on faith. That contract is the wrong default precisely where Round Table is empirically weakest: the Echo Chamber Limit (Section 6) and the absent same-provider control for the Editorial Swarm error-catch (Section 5.1.2) both call for re-adjudicating a past debate with a disjoint judge, and an irreversibly discarded trace cannot be re-adjudicated. We therefore adopt the reversible Compress-Cache-Retrieve contract that headroom (Chopra, 2026) establishes for generic agent context: the propose/challenge/defend exchange is compressed out of the active window but cached behind a retrieval pointer, so a later debate round, a repair cycle, or an audit can recover the full exchange that produced a conclusion instead of working only from the synthesis. The two systems are complementary, not redundant—headroom's compressors are content-type-aware and deliberation-blind, while Round Table's compaction is schema-aware but, until this change, one-directional. Combining them yields a deliberation-schema-aware compaction layer over a reversible CCR backing store: the synthesis remains the cheap default view, and the originals remain recoverable.

Measuring what compaction costs. The paper previously asserted that compaction "discards redundant intermediate exchanges" without measuring the fidelity lost in the process—a gap headroom's reproducible token-savings-versus-accuracy harness throws into relief. We adopt the same discipline as a required, not optional, evaluation: the compaction step must be characterized by a token-savings-versus-fidelity curve, where fidelity is the rate at which a

disjoint judge, given only the compacted synthesis-plus-dissents, reaches the same ratification verdict it reaches with the full retrievable trace. We flag this compaction-fidelity eval as concrete future work rather than a shipped measurement; the reversible cache is the precondition that makes it runnable, because it preserves the ground-truth trace the judge needs to compare against.

This persistence also turns deliberation into a multi-turn process rather than a one-shot. The implementation retains per-debate speaker memory (Section 11.3), so the same model set can resume from its last synthesis instead of re-arguing every phase from zero—closer to the human habit of "sleeping on it" and revisiting a decision with the prior reasoning already in hand. A nightly reflection cycle in the host agent can also identify decisions that would benefit from multi-model deliberation and schedule them automatically—in effect, the agent learns *when* to think harder. Such a reflective scheduler—a background process that reviews the day's decisions and queues the hard ones for deeper treatment—is described as a companion mechanism in separate work (Serra, forthcoming).

5.1 Editorial Swarm: Parallelized Multi-Artifact Review (Preliminary Deployment Observation)

The preceding sections describe Round Table as a protocol for deepening deliberation on a single question. In production, we encountered a complementary need: applying Round Table principles to *N independent artifacts simultaneously*. This section documents the **Editorial Swarm** pattern, deployed in March 2026 for a batch review of eight research papers. We present this as a preliminary deployment observation (N=1 deployment, no control group) rather than a controlled experiment, and discuss its limitations explicitly.

5.1.1 Deployment Architecture

The Editorial Swarm processed eight papers through three rounds of parallelized cross-provider review, orchestrated by a single Claude Opus 4 instance acting as coordinator.

Round 1 — Cross-Provider Parallel Edit. Eight sub-agents were spawned simultaneously: four Claude Opus 4 instances and four GPT-5.4 instances, each assigned one paper. Cross-judging was enforced: GPT instances reviewed papers edited by Opus, and vice versa. This cross-provider constraint was designed to maximize the probability of surfacing errors that a same-provider review would miss.

Round 2 — Adversarial Bounce Cycles. Eight Opus editors were spawned, each internally invoking GPT-5.4 as a harsh critic. Each editor–critic pair engaged in 2–3 bounce rounds targeting a convergence threshold of 9.8/10 on a structured quality rubric. The bounce rounds implemented a micro-RAAC cycle: propose (Opus edit) → challenge (GPT critique) → defend (Opus revision) → ratify (GPT re-score).

Round 3 — Convergence Enforcement. The same bounce pattern repeated with an explicit convergence target. Papers that had not reached the 9.8/10 threshold were re-spawned with augmented critique instructions.

At peak concurrency, up to **24 agents** operated simultaneously across two provider subscriptions (Anthropic and OpenAI), with the orchestrator tracking convergence state for all eight papers.

5.1.2 Cross-Provider Error Detection: An Illustrative Case

During Round 1, a GPT-5.4 reviewer identified a misleading statistic in an Opus-edited paper: the paper reported 93% accuracy on a classification task, but the underlying data showed 79% overall accuracy with 5 of 6 classes at zero performance—the 93% figure reflected only the majority class. Opus had missed this in its editorial pass.

This incident is consistent with the core Round Table thesis—that cross-provider review can catch errors that same-provider review might miss—but we are careful not to overstate a single case. We did not run a same-provider control (e.g., an Opus-reviewing-Opus baseline), and therefore cannot definitively attribute the catch to cross-provider diversity rather than to reviewer assignment variance. The incident is illustrative, not conclusive.

5.1.3 Observed Model Role Tendencies

Production deployment suggested the following role tendencies, which are operational observations from a single deployment rather than empirically validated characterizations:

Model	Observed Strength	Observed Weakness	Suggested Role
Claude Opus 4	Coherence across long documents; structural reasoning; synthesis	Accepts plausible-sounding claims without statistical verification	Coordinator, Editor, Synthesizer
GPT-5.4	Finding logical gaps, statistical errors, weak claims; compressing critique	Compresses and summarizes when generating long documents (Section 4.1)	Critic, Reviewer
Gemini 2.5 Pro	Pushing harder on assumptions; finding what agreeable reviewers miss	Less consistent on structured multi-step editorial tasks	Devil's Advocate
Manus	Consolidating review notes; formatting; mechanical tasks	Limited reasoning depth on substantive content	Secretary, Consolidator

5.1.4 Depth vs. Breadth: Scaling Dimensions

The Editorial Swarm suggests that Round Table principles operate along two scaling dimensions:

- **Depth scaling** (original RAAC): More rounds of debate on a single question, deepening analysis. Cost scales as $O(n^2 \cdot T_{\max})$.
- **Breadth scaling** (Editorial Swarm): More parallel instances of the protocol across N independent artifacts. Cost scales as $O(N \cdot n^2 \cdot T_{\max})$ but *parallelizes* across providers and subscriptions.

These axes appear orthogonal in principle: an Editorial Swarm can use shallow debate (1-round Tribunal per artifact) or deep debate (5-round Full Round Table per artifact), depending on the stakes and complexity of each artifact. We have not yet empirically characterized the marginal returns of depth vs. breadth allocation, which would require systematic experimentation across task types and complexity levels.

5.1.5 Cost and Operational Summary

Metric	Value
Papers reviewed	8
Rounds	3
Agents per round (avg)	~2 per paper
Total agent-turns	~48

Metric	Value
Provider subscriptions	2 (Anthropic, OpenAI)
Peak concurrent agents	24

Running across two provider subscriptions yielded two operational benefits: (1) **rate-limit parallelism**—neither provider's rate limits bottlenecked the pipeline; and (2) **failure isolation**—a provider outage would degrade throughput by ~50% rather than halting the pipeline entirely.

5.1.6 Limitations of the Editorial Swarm Observation

This deployment has significant limitations as evidence:

1. **No control group.** We did not run a same-provider baseline (e.g., Opus-only swarm), so we cannot isolate the contribution of cross-provider diversity from other factors (reviewer quality, task assignment, prompt design).
2. **N=1 deployment.** A single deployment of 8 papers does not constitute a reproducible experiment. The observations may not generalize to other domains, artifact types, or model versions.
3. **No quantitative quality metrics.** We did not measure inter-rater agreement, error detection rates, or revision quality using standardized metrics. Quality assessment was based on the orchestrator's rubric-based scoring, which is subject to the same echo-chamber concerns discussed in Section 6.
4. **Survivorship bias.** We report errors that were caught; we do not know the population of errors that were missed. Without a ground-truth error inventory, the error detection rate is unknown.

Despite these limitations, we include the Editorial Swarm because it demonstrates the *feasibility* of breadth-scaled cross-provider deliberation in production and generated operational hypotheses (compression tendency, role specialization) that warrant controlled investigation.

6. Limits of Deliberation

The Echo Chamber Limit. An ensemble cannot validly evaluate its own output. Models grading their own debate syntheses exhibit sycophantic agreement, inflating scores. Evaluation therefore requires disjoint models or external ground-truth oracles. Our consensus quality metric (ratification votes from debate participants) is subject to this limit, and external evaluation would provide stronger evidence.

The Single-Agent Self-Debate Question. A legitimate challenge to multi-agent frameworks is that a single frontier model, prompted to argue from multiple perspectives, might achieve equivalent variance reduction at lower cost. Wang, Z. et al. (2025) find that multi-agent debate outperforms single-agent self-consistency specifically under conditions of high inter-agent diversity—precisely the regime CDI measures. Li et al. (2025) find that single-agent approaches can match multi-agent setups on some tasks, but that genuine cross-provider diversity provides irreducible benefits when tasks require diverse domain expertise. Round Table's operating regime is therefore explicitly the high-CDI regime: when providers have genuinely different training-induced error profiles, multi-agent debate extracts value that single-agent self-debate cannot. For homogeneous ensembles (low CDI), single-agent approaches may be preferable. We have not yet run a controlled single-agent self-debate baseline against Round Table, which we identify as a priority for future work.

The strongest such baseline is now concrete rather than hypothetical: addyosmani/agent-skills ships **doubt-driven-development** (Osmani, 2026), a deployed single-provider adversarial protocol that runs CLAIM → EXTRACT (artifact plus contract, with the original reasoning chain stripped) → DOUBT (a *fresh-context* adversarial reviewer that has no access to the producing chain of thought) → RECONCILE (each finding classified against the artifact text) → STOP (terminating on a trivial-findings test, a 3-cycle cap, or explicit user override, with stated non-trivial-decision criteria). doubt-driven-development is the right competitor to name because it manufactures its review diversity through *context isolation*—a reasoning-stripped, fresh-context reviewer—rather than through cross-provider error-decorrelation. The honest comparison cuts both ways. What Round Table has that doubt-driven-development does not: measured CDI showing the providers' errors are genuinely decorrelated ($\Sigma \approx 0.38$), so the disagreement is not merely an artifact of one model's sampling variance, plus a ratification vote that structurally defends against the Persuasion Paradox. What doubt-driven-development has that RAAC lacks: an explicit, reasoning-stripped artifact handoff that prevents the reviewer from anchoring on the author's chain of thought, and a principled STOP criterion (trivial-findings plus cycle-cap) in place of RAAC's pure round-cap—which the paper admits never reaches formal convergence (Section 6, "The Convergence Gap"). We therefore identify two concrete actions: (a) run the missing single-agent ablation with doubt-driven-development as the named single-provider baseline, holding the task set fixed; and (b) adopt its EXTRACT and STOP discipline into RAAC—handing each Challenge-phase reviewer a reasoning-stripped position and adding a trivial-findings/cycle-cap stop test alongside the round-cap—so that convergence is governed by whether new findings are still material, not only by exhausting a fixed round budget.

The Orchestration and Latency Limit. Synchronous N-model debate hits API rate limits quickly. Chen, T. et al. (2026) measure multi-agent orchestration overhead at over 100× single-agent latency in some configurations. Round Table is fundamentally an asynchronous, offline protocol best suited for tasks where deliberation quality justifies minutes-scale latency. Real-time applications require the lighter Tribunal or Fan-Out architectures. The five-phase protocol is also inherently sequential—each phase waits on the slowest participant of the prior phase—which bounds how fast a full debate can complete. A streaming variant, in which ratification votes are admitted as each participant completes rather than gated on a synchronous barrier, is a natural way to cut wall-clock latency without changing the protocol's semantics; we flag it as concrete future work rather than a shipped capability. The current benchmark harness uses synthetic mock participants for reproducible validation; the production path described in Section 11 runs against live cross-provider models through an asynchronous spawn-and-wait orchestration layer.

The Supply Feasibility Limit. *(v8.6, 2026-09-04.)* The protocol assumes that when a debate is warranted, the panel can be *convened* — that three houses will answer when called. On a harness running against metered API keys that assumption is safe. On one running against **subscriptions** it is not, and the failure is not graceful: a panellist that is rate-limited does not deliberate worse, it does not deliberate at all, and a five-phase protocol that loses a participant between Challenge and Ratification has silently become a different, smaller ensemble than the one whose CDI was measured. Section 11.3's dropout recovery keeps such a debate from being *poisoned*; it cannot keep it from being *less diverse than it was designed to be*, which is the property the whole paper rests on.

Measured instance. On 2026-09-03 at 15:39 UTC, a live three-supply board on our own harness read: Claude 71% of its weekly window, ChatGPT **100%** of its 5-hour and 92% of its weekly, GitHub Copilot **100%** of its monthly premium allowance, and xAI 28% of a weekly-only window. A three-house panel convened that afternoon would have been a one-house panel with two silences, and CDI would have reported the diversity of the ensemble that was *requested* rather than the one that *ran*.

Two consequences follow, and the second is the one we did not anticipate.

First, **a panel needs a feasibility precondition alongside its diversity precondition.** CDI answers "are these models decorrelated?"; it does not answer "can these models be reached, and can the portfolio absorb n simultaneous calls?" Both must hold, and only the first is currently formalised here. The practical form is a veto evaluated before role assignment: a participant whose supply is spent, cooling, or unfunded is not a low-affinity participant, it is an absent one, and it should be replaced during matching rather than discovered missing at Challenge.

Second, **provider diversity and supply diversity are different quantities, and the protocol prices one while paying for the other.** CDI deduplicates by *provider*, on the sound theory that shared training lineage implies shared blind spots. But the cost of a panel is paid per *billing pool*, and the two do not partition the same way: two distinct providers can draw on one subscription (so a "two-house panel" is one bucket's worth of exposure), and one house can be reachable through two independent pools (so a same-provider pair may be, financially, two panellists). A panel selected for maximal CDI can therefore be maximally concentrated in supply — the configuration most likely to lose two participants to the same limiter at the same moment. We flag this as an unmeasured confound in the deployed ensemble rather than a corrected one; the fix is to require distinctness on *both* axes during bipartite matching, and we have not yet re-run the benchmark under that constraint.

This limit is developed as a general mechanism in the companion allocation paper (Serra, 2026, *MAESTRO* §4.5, S1 and S5), where the same observation generalises beyond debate: capability mechanisms buy quality with extra calls, and on a clock-metered portfolio each extra call is extra exposure to the limiter that ends the turn. Debate is simply the mechanism that buys the most quality per unit of that exposure, which makes it both the best case for the argument and the most sensitive to it.

The Convergence Gap. No scenario in the benchmark reached formal convergence (the text-stability-plus-acceptance threshold) within the capped round limit, though all reached ratification convergence. This is structurally expected with aggressive round caps; deeper production tiers should converge more reliably, though formal guarantees remain open.

The Perspective Coverage Ceiling. Keyword-matching evaluation yields average coverage of 0.10 (10%). Synthesis outputs integrate expected perspectives *semantically*, but not through exact keyword strings. Future evaluation should use embedding-based similarity.

Affinity Matrix Brittleness. The role affinity scores reflect current training methodologies. As providers update their training pipelines, optimal role assignments may shift, and the matrix should be periodically recalibrated. The **Parity Assumption** underlying the Diversity Premium (Section 7.3) also deserves scrutiny: if one provider achieves a structural capability leap, forcing a weaker model into the ensemble could be actively harmful. CDI is a necessary but not sufficient condition for ensemble benefit; minimum per-model competence is also required.

Persuasion Robustness Is Asserted, Not Yet Measured. The Ratification phase is designed to neutralize the Persuasion Paradox (Khan et al., 2024; Wan et al., 2024) by giving every participant an equal final vote regardless of how persuasively the synthesis is argued. The deployed protocol additionally prevents a failed or non-responding participant from being counted as agreement (Section 11.3). But we have not yet run the adversarial stress test this claim warrants: deliberately seeding one model with a confident-but-wrong position and measuring whether it can drag the ensemble to an incorrect ratified consensus, across role assignments and ensemble sizes. Until that test is run, persuasion robustness is a design intent supported by the protocol's structure, not an empirically demonstrated property.

Missing Ablations. A full ablation studying CDI vs. accuracy across ensemble sizes (2, 3, 4, 5 models) and domain types would strengthen the empirical case. We also lack direct comparison against MoA (Wang, Y. et al., 2024a) on matched benchmarks, which is needed to quantify the advantage of adversarial over cooperative aggregation. Hypothesis H4 (ECE reduction) remains

untested. A controlled single-agent self-debate baseline—run against doubt-driven-development (Osmani, 2026) as the named single-provider competitor—is needed to isolate the value of cross-provider diversity from the value of structured adversarial deliberation itself.

Single-Run Variance. GPQA Diamond results are from a single run. While the 8-point gain is large relative to expected inter-run variance, multi-run confidence intervals are needed for definitive claims.

7. The Diversity Premium

7.1 Value of Error Reduction

Let V_{err} be the cost of an error, C_{debate} the cost of debate (including both token costs and latency), and ΔE the error reduction. Debate is net-positive when $\Delta E \cdot V_{\text{err}} > C_{\text{debate}}$.

7.2 Break-Even Analysis

Error Cost (V_{err})	Debate Cost (tokens)	Typical Latency	Required ΔE	Typical Domain
\$50	\$0.01 (Tribunal)	~5s	0.02%	Content generation
\$200	\$0.08 (Full RT)	~30–120s	0.04%	Software engineering
\$5,000	\$0.08 (Full RT)	~30–120s	0.002%	Medical decision support
\$50,000	\$0.08 (Full RT)	~30–120s	0.0002%	Legal/regulatory compliance

Even at the Full Round Table cost tier, the break-even error reduction is negligible for any domain where errors have material consequences. The latency column makes a critical trade-off explicit: Full Round Table debate requires 30–120 seconds per query depending on model response times and round count (consistent with the 100×+ overhead measured by Chen, T. et al., 2026 for multi-agent orchestration). This makes Full Round Table unsuitable for latency-sensitive applications (real-time chat, interactive coding), but appropriate for batch processing, document review, critical decision support, and other offline tasks where deliberation quality justifies the time cost. The Tribunal pattern, at ~5s latency, is viable for near-interactive use cases.

Caveat. Token costs and latency estimates are based on current API pricing (2026) and our deployment measurements. Both vary significantly by provider, model version, and task complexity. The required ΔE calculations depend on assumed error costs that are domain-specific and not empirically derived in this work.

7.3 Diversity as Positive Externality

In a market with one dominant provider, CDI $\to 0$. In a market with competing providers with distinct training approaches, CDI > 0. The existence of diverse architectures creates a **Diversity Premium**: users who combine models can outperform users of any single model. This reframes model commoditization as a driver of systemic reliability rather than a threat. The premium is largest when providers maintain genuinely different training methodologies rather than converging on identical approaches.

Fragility of the Diversity Premium. The premium depends on approximate parity among frontier models. If one provider achieves dominant accuracy across all domains, the CDI of any mixed ensemble may decrease, and the weaker models may introduce errors rather than correct them. The Diversity Premium is therefore a structural property of the current multi-provider landscape, not a guaranteed permanent advantage.

8. Evaluation Design

8.1 Research Hypotheses

ID	Hypothesis	Metric	Status
H1	High-CDI heterogeneous debate outperforms low-CDI homogeneous debate	Accuracy ($p < 0.05$)	Supported (GPQA)
H2	Performance gain correlates positively with CDI	Pearson $r > 0.7$	Untested (requires multi-ensemble ablation)
H3	Optimal RAAC assignment outperforms random role assignment	Win rate > 60%	Supported (73%, p < 0.01)
H4	Debates produce lower Expected Calibration Error (ECE) than single models	ECE decrease	Untested

We are explicit about which hypotheses have been tested and which remain open. H2 in particular requires a systematic ablation across ensemble configurations that we have not yet conducted.

8.2 Setup and Baselines

Benchmarks: GPQA Diamond (Science), MATH-500, HumanEval, TruthfulQA. **Baselines:** Single Model Best-of-1, Self-Consistency Best-of-5, Homogeneous Debate (Du et al., 2023), MoA (Wang, Y. et al., 2024a). **Round Table Configuration:** Claude Opus (Architect), GPT-o3 (Critic), Gemini 2.5 Pro (Pragmatist).

Note on baselines. We report results against Single Model Best-of-1 (Section 9.2). Comparison against Homogeneous Debate and MoA on matched benchmarks has not been completed and is a priority for future work; we therefore cannot yet claim that adversarial cross-provider debate outperforms cooperative aggregation or same-provider debate.

8.3 Power Analysis

To detect a medium effect ($d = 0.5, \alpha = 0.05, \beta = 0.20$), required $N = 64$ for a two-sample proportion test. Our smallest set (HumanEval, $N = 164$) provides power > 0.99. Applying this generic power analysis to accuracy differences on discrete benchmark tasks involves simplifying assumptions about effect size and distributional properties.

9. Infrastructure and Results

9.1 Test Infrastructure

Testing uses an offline arena with sandbox isolation. Participants are synthetic mock models with deterministic, role-differentiated response generators. Ground-truth evaluation uses parsed oracles (Python execution, SymPy). A trace recorder logs all debate events for offline CDI and cost analysis. The arena requires no live API calls, enabling fully reproducible benchmark execution in CI.

Important methodological note: The GPQA Diamond accuracy figures (Section 9.2) were obtained by running the 3-model ensemble against live API endpoints (temperature = 0.3, greedy decoding for baselines, single run per question, $n = 198$). The protocol validation metrics (CDI, consensus quality) in Sections 9.3–9.5 used synthetic participants. We distinguish these clearly because the validity claims differ: live results demonstrate real accuracy gains; synthetic results validate protocol mechanics (state machine correctness, convergence behavior, cost accounting) but *cannot* validate claims about real cross-provider cognitive biases or the value of role amplification. Claims about cognitive diversity and role assignment benefits rest on the live GPQA results and the role assignment sensitivity analysis, not on the synthetic protocol validation.

9.2 Accuracy Gains (Live Models)

The 3-model Round Table ensemble achieves **63.6%** on GPQA Diamond ($n = 198$) vs. Claude Opus 54.0%, GPT-o3 55.6%, Gemini 2.5 Pro 53.8%. This is an **8 percentage-point absolute gain** (14.4% relative) over the strongest single model. On HumanEval pass@1 ($n = 164$): 46.3% vs. 41.2%, 39.0%, 38.5% individually. Adversarial ratification caught edge-case reasoning failures that individual models missed.

Caveats. These are single-run results. The 8-point GPQA gain, while large relative to expected inter-run variance on a 198-question benchmark, has not been confirmed across multiple runs with different random seeds. We have not computed confidence intervals on the accuracy difference or performed a formal significance test (e.g., McNemar's test on paired per-question outcomes). Multi-run replication is a priority for future work.

9.3 CDI Measurements

Factual benchmark (GPQA): Pairwise Pearson correlation on greedy-decoded error vectors yields average $\Sigma_{ij} \approx 0.38$, giving **CDI $\approx$ 0.62** ($k = 198$). This indicates substantial independence across frontier model failure modes.

Open-ended design benchmark (5 participants, 5 scenarios, synthetic): CDI = **1.06**. Given the small scenario count ($k = 5$), this point estimate carries high uncertainty and the previously reported 95% CI ([0.60, 1.49]) should be treated as approximate due to the statistical limitations discussed in Section 3.1. The elevated CDI relative to GPQA (0.62) may reflect broader solution spaces where role amplification drives more divergent approaches, but could also be an artifact of the small sample or synthetic participant design.

9.4 Cost Analysis

Full Round Table debate consumes ~15× the token volume of single zero-shot. At that multiplier, Round Table yields an 18% relative error reduction on challenging benchmarks.

Benchmark cost measurements (5 scenarios, synthetic participants):

Architecture	Avg Cost/Scenario	Range
Full Round Table (2 rounds)	$0.0842	$0.0839–$0.0845

Architecture	Avg Cost/Scenario	Range
Moderated Tribunal (1 round)	$0.0101	$0.0093–$0.0118
Cost ratio	**8.3**×	—

The 8.3× ratio matches the theoretical $O(n^2)$ vs. $O(n{+}1)$ scaling prediction. Total benchmark cost: $0.20. These synthetic-participant figures fix the protocol's *structural* cost shape; absolute per-debate cost on live cross-provider models is governed by the depth tier's budget ceiling (Section 11.3) and remains to be characterized against a production cost log, which we identify as the next measurement to close.

9.5 Protocol Validation (Synthetic)

Scenario	Domain	Architecture	Rounds	Consensus Quality	Cost (USD)
s1-arch	Software Architecture	full-round-table	2	0.80	$0.0839
s2-ai-safety	AI Ethics	full-round-table	2	0.80	$0.0845
s3-data	ML Systems	moderated-tribunal	1	0.80	$0.0093
s4-product	Product Strategy	moderated-tribunal	1	0.80	$0.0093
s5-infra	Infrastructure	moderated-tribunal	1	0.80	$0.0118

Consensus quality 0.80 = 4 of 5 participants accepting in final ratification. The uniform 0.80 across all scenarios reflects the deterministic nature of synthetic participants rather than a robust quality signal. These results validate protocol *mechanics* (state transitions, cost accounting, convergence checks) but should not be read as evidence of deliberation *quality*. Live-model protocol validation is needed to assess quality variation across domains.

9.6 Automated Test Suite

The implementation is validated by an automated test suite (vitest) spanning CDI computation, RAAC role-assignment and bipartite matching, the 5-phase debate state machine, ratification voting and repair cycles, persistent-deliberation compaction, and the architecture patterns. Subsequent development added test coverage for the diversity-lock, dropout-recovery, cost-budget, speaker-selection, and orchestrator mechanisms described in Section 11.3, growing the suite to roughly 110 cases across nine test files as those capabilities landed. We report the suite as a correctness gate for protocol mechanics, not as evidence of deliberation quality—the latter requires live-model evaluation (Section 9.5).

10. Role Assignment Sensitivity

To validate H3 (optimal vs. random role assignment), we compared debate quality under three conditions: (1) optimal RAAC matching using the affinity matrix, (2) random role assignment, and (3) adversarial misassignment (each model assigned its lowest-affinity role). Optimal assignment outperforms random in 73% of trials (N=30, $p < 0.01$) as measured by ratification acceptance rate. Adversarial misassignment degrades acceptance rate by 12–18% versus optimal, confirming that the affinity structure captures meaningful signal.

Limitations of this analysis: Synthesis quality was evaluated by ratification votes from debate participants, which is subject to the Echo Chamber Limit (Section 6). The acceptance rate measures internal consensus, not objective output quality. An external evaluation—using a disjoint judge model or human assessment—would provide stronger evidence. Additionally, since the affinity matrix was derived in part from pilot task observations, there is a risk of circularity: the matrix may encode the same patterns it is being evaluated against. We leave independent external evaluation to future work.

11. Implementation

11.1 Module Structure

Round Table is implemented as a self-contained extension to a persistent-agent runtime (public release planned upon acceptance). The codebase is organized around a small number of single-responsibility modules:

- **cognitive-diversity** — CDI measurement, Fisher z confidence intervals, provider-diversity selection.
- **raac-protocol** — the RAAC 5-phase protocol, role affinity, convergence, budget gate, and dropout recovery.
- **debate-architectures** — Fan-Out, Moderated Tribunal, Full Round Table, and Tournament patterns.
- **real-participant** — role → cross-provider model routing for live debate.
- **speaker-selection-api** — a pluggable speaker / role-assignment hook.
- **orchestrator-api** — the swappable debate-orchestrator contract, with RAAC as the default.
- **persistent-deliberation** — durable artifact storage, the deliberation-memory schema, and speaker memory.

These modules are exercised by an automated test suite spanning nine test files. The extension is TypeScript ESM on a current Node runtime. The agent-facing surface is a single deliberation tool that takes a topic plus options (debate depth, optional resume handle) and returns the consensus, a confidence estimate, recorded dissents, action items, and the ensemble's diversity score.

11.2 Architecture Notes

CDI Pipeline. Computes the Phi coefficient between binary error vectors, aggregates pairwise correlations, and returns CDI with an approximate 95% CI via Fisher z-transform (SE = $1/\sqrt{k-3}$, z-critical = 1.96). As discussed in Section 3.1, these CIs are approximate due to non-independence of pairwise correlations.

RAAC Protocol. A single orchestration routine drives the 5-phase loop with the deployed defaults: convergence weight $\lambda = 0.5$, convergence threshold $\epsilon = 0.1$, ratification threshold 0.5, and per-depth round caps of 2 / 4 / 6 for the `quick` / `standard` / `deep` tiers. Role assignment uses greedy maximum-weight bipartite matching on the affinity matrix. Greedy matching is not guaranteed optimal for all inputs; the Hungarian algorithm would provide exact solutions. In practice, with $n \leq 5$ models and well-separated affinity scores, greedy matching produced optimal assignments in all tested configurations. The text-stability term of the convergence test is computed as Jaccard overlap of consecutive syntheses' word sets—a lightweight, dependency-free proxy for embedding distance (Section 4.2)—rather than a learned-embedding cosine.

Debate Architectures. Four architecture functions, each returning a typed result with a per-phase cost breakdown.

Persistent Deliberation. A session-scoped deliberation manager stores structured JSON artifacts (Appendix B) as durable memory events. On session start it loads the latest deliberation memory to resume without replaying raw traces, and it retains per-debate speaker memory so a follow-up call with the same handle resumes from the prior synthesis (Section 5).

11.3 From "We Hope They're Diverse" to "We Guarantee It": Diversity, Robustness, and Orchestration

The cross-provider thesis is only credible if the running system actually maintains cross-provider diversity under real-world configuration, failure, and cost pressure. Earlier iterations relied on a fixed five-model loop with hardcoded provider references; the current implementation replaces that with a set of mechanisms that make the diversity claim *enforceable* and the protocol *open*. We summarize them as deployed capabilities, and are explicit about which parts are fully active versus gated on infrastructure that has not yet shipped.

Configurable cross-provider routing. Role-to-model mapping is no longer baked into the protocol. Each role resolves to a model reference through a precedence chain—per-deployment override, then a built-in default, then a final fallback—so a deployment whose catalog lacks a given provider keeps genuine cross-provider diversity instead of silently collapsing every role onto one fallback model. Every substitution emits a warning at debate start, so a diversity collapse is visible rather than silent. (As noted in Section 4.1, our reference deployment has no dedicated deep-exploration model configured and therefore routes the Researcher role to an alternative high-reasoning model, yielding a three-vendor spread rather than the four-archetype matrix.)

Provider-diversity lock. When enabled, participant selection guarantees at most one participant per *resolved* provider: vendor identity is derived from the resolved model reference, not from a cosmetic label, and duplicates are dropped greedily by lowest affinity. Because the lock operates on the selected set, a strict lock can shrink the ensemble to the number of distinct available providers—an honest trade we surface rather than hide (with only three distinct vendors available, a strict five-participant debate is not achievable). This turns "we hope the participants are diverse" into "we guarantee at most one per provider."

Dropout recovery. A participant that times out or errors previously emitted a sentinel string that then flowed into the synthesis as if it were a genuine position—poisoning the consensus and, worse, silently shrinking the diversity that the ratification safeguard depends on. The implementation now detects that sentinel, promotes a backup participant from a not-yet-represented provider, and—if the backup also fails—records the dropout and proceeds with the remaining participants rather than treating absence as agreement. This directly defends the protocol's robustness claim against the single-dominant-voice failure mode, though as Section 6 notes the adversarial measurement of that robustness is still future work.

Cost-aware budget gate. Each depth tier carries a USD budget ceiling (1 / 3 / 8 for quick / standard / deep), and the debate loop halts when spending reaches the cap. Above that per-debate ceiling the implementation clamps the budget to a fraction of the agent's real billing headroom, so a single deep debate cannot exhaust a small metered budget. We are precise about its status, which has changed since this layer was first described: the host runtime now performs cost-aware model-and-effort gating on every subagent dispatch—each fanned-out debate participant is coerced to a cost-tier model and the billing-headroom signal the clamp was waiting on is consumed at the dispatch boundary—so the outer clamp is live rather than the verified no-op it once was. We keep the scientific caution this warrants: there is still no end-to-end production cost log validating the gate's behavior under real billing pressure (Section 9.4), so what is demonstrated is that the dispatch-layer signal is now wired and enforced, not yet that the clamp behaves correctly across a full production cost distribution. The cost ceiling is enforced outside any individual orchestrator (see below) so it survives orchestrator swaps.

Pluggable speaker selection and swappable orchestration. Two interfaces make the deliberation substrate open rather than fixed. A speaker-selection hook lets an external manager—or a different cognitive architecture—decide who speaks in which role, with the built-in greedy match as the default and automatic fallback. A debate-orchestrator contract lets an entirely different choreography drive the same cross-provider call fabric: the RAAC 5-phase protocol is the default orchestrator, the existing architecture functions (fan-out, sequential, moderated-tribunal) are exposed as alternative built-in orchestrators, and an external orchestrator can be supplied by a sibling extension. This is the step that turns Round Table from "a protocol" into "a platform"—the strongest form of the open-substrate claim. The chief residual risks are honest ones: non-RAAC orchestrators produce sparser traces than the round/ratification shape the persistence layer expects, and an external orchestrator that ignores the budget ceiling would defeat the cost gate unless the cap is enforced around it (which it is).

A note on the right activation trigger. Multi-model debate is expensive, so the open question is less "can the agent debate?" than "when should it?" A persistent agent's reflective cycle, or a designated high-stakes step in a multi-step task recipe, is a more principled activation point than relying on the agent to remember to invoke deliberation. Wiring the debate tool as the body of such a step gives Round Table a built-in "think harder here" hook; we treat this integration as a deployment pattern rather than a protocol feature.

11.4 Host-Runtime Integration Points

Round Table is designed to sit inside a larger persistent-agent runtime, and two host capabilities sharpen its behavior. The first is the durable event store: deliberation traces are written to the host's episodic memory and compacted into consensus-plus-dissent summaries, the mechanism described in Section 5. The compaction layer is an episodic-memory system that stores interactions as typed events and later distills them into summaries that retain the conclusion and key dissents while discarding redundant intermediate exchanges (Serra, forthcoming). The second is persona-aware context engineering: the host agent's configured persona governs debate-depth selection, so a persona tuned for technical rigor selects different model combinations and deeper tiers than one tuned for creative brainstorming. A persona here is a persisted set of preferences and reasoning dispositions that the agent carries across sessions (Serra, forthcoming). Neither capability is required to run Round Table—it degrades to a stateless single-shot deliberation tool without them—but both extend it from a protocol into a standing reasoning faculty of the agent.

12. Conclusion

Round Table converts model heterogeneity from a source of friction into a mechanism for reliability. The core contribution is matching each model to the role its training-induced cognitive tendencies make it best suited for, then structuring adversarial debate to amplify the productive tension between complementary cognitive styles—a board of directors where each member has different expertise and different blind spots, and where the synthesizer does not average the opinions but identifies where they agree (a strong signal) and where they disagree (the areas that need deeper investigation). CDI provides a quantitative measure of ensemble diversity; RAAC provides the deliberation protocol; Persistent Deliberation extends single-shot debate into multi-session reasoning.

Empirical results confirm an 8 percentage-point accuracy gain on GPQA Diamond (single run), CDI = 0.62 for the three-model GPQA ensemble and 1.06 for the five-model open-ended panel (small k), and consensus quality of 0.80 on synthetic protocol validation. The Editorial Swarm deployment provides preliminary observational evidence that cross-provider deliberation

catches reasoning-level errors and that models exhibit role-specific strengths consistent with their training methodologies. Beyond these results, the deployed implementation makes the cross-provider claim enforceable: configurable per-role routing, a provider-diversity lock, dropout recovery, a cost-aware budget gate, and swappable speaker-selection and orchestration interfaces together turn "we hope the ensemble is diverse" into "we guarantee it, and we can prove it from the resolved provider mix."

Significant open questions remain: multi-run replication of the GPQA result, systematic testing of the CDI–performance relationship (VR-1), a controlled single-agent self-debate baseline, comparison against cooperative aggregation, empirical grounding of the affinity matrix, an adversarial persuasion-robustness stress test, and a production cost log validating the now-live budget gate against a real billing distribution. We view these not as weaknesses to apologize for but as a research agenda motivated by a framework that is already producing measurable gains in production.

As individual model performance approaches asymptotic limits, the next frontier is orchestrating cognitively diverse ensembles where each model is placed at its point of greatest comparative advantage.

References

- Anthropic (2024). *The Claude 3 Model Family: Opus, Sonnet, Haiku.*
- Brown, T. et al. (2020). *Language Models are Few-Shot Learners.* NeurIPS.
- Burns, C. et al. (2023). *Weak-to-Strong Generalization: Eliciting Strong Capabilities With Weak Supervision.* OpenAI.
- Chen, L. et al. (2023). *FrugalGPT: How to Use Large Language Models While Reducing Cost and Improving Performance.*
- Chen, T. et al. (2026). *Understanding Multi-Agent LLM Frameworks: A Unified Benchmark and Experimental Analysis.* arXiv:2602.03128.
- Chopra, T. (2026). *headroom: Reversible Compress-Cache-Retrieve Context Compression for LLM Agents.* GitHub: chopratejas/headroom.
- Condorcet, N. (1785). *Essay on the Application of Analysis to the Probability of Decisions.*
- DeepSeek-AI (2025). *DeepSeek R1: Reinforcement Learning for Reasoning.*
- Dietterich, T. G. (2000). *Ensemble Methods in Machine Learning.*
- Du, Y. et al. (2023). *Improving Factuality and Reasoning in Language Models through Multiagent Debate.* ICML.
- Feng, L. et al. (2026). *Dr. MAS: Stable Reinforcement Learning for Multi-Agent LLM Systems.* arXiv:2602.08847.
- Hansen, L. K., & Salamon, P. (1990). *Neural Network Ensembles.* IEEE TPAMI.
- Hao, Y. et al. (2025). *Understanding LLM Behaviors via Compression: Data Generation, Knowledge Acquisition and Scaling Laws.* arXiv:2504.09597.
- Hong, S. et al. (2024). *MetaGPT: Meta Programming for A Multi-Agent Collaborative Framework.* ICLR.
- Irving, G. et al. (2018). *AI Safety via Debate.*
- Journey (2026). *Journey Kit Registry: Installable Agent Workflows.* journey registry.
- Jiang, D. et al. (2023). *LLM-Blender: Ensembling Large Language Models with Pairwise Ranking and Generative Fusion.* ACL.
- Khan, A. et al. (2024). *Debating with More Persuasive LLMs Leads to More Truthful Answers.* ICML.
- Krogh, A., & Vedelsby, J. (1995). *Neural Network Ensembles, Cross Validation, and Active Learning.* NIPS.
- Kuncheva, L. I., & Whitaker, C. J. (2003). *Measures of Diversity in Classifier Ensembles*

and Their Relationship with the Ensemble Accuracy. Machine Learning, 51(2).

- Li, X. et al. (2025). *Single-agent or Multi-agent Systems? Why Not Both?* arXiv:2505.18286.
- Margineantu, D. D., & Dietterich, T. G. (1997). *Pruning Adaptive Boosting.* ICML.
- Osmani, A. (2026). *agent-skills: A Discipline-Oriented Agent Skill Library (incl. doubt-driven-development).* GitHub: addyosmani/agent-skills.
- Serra, O. (forthcoming). *Total Recall: Event-Navigated Graded Retrieval & Archival Memory.*
- Serra, O. (forthcoming). *Identity Persistence: Persona-Aware Context Engineering for Persistent AI Identity.*
- Silva, A. et al. (2025). *A Taxonomy of Hierarchical Multi-Agent Systems: Design Patterns, Coordination Mechanisms, and Industrial Applications.* arXiv:2508.12683.
- Wan, X. et al. (2024). *The Persuasion Paradox: When Confidence Mimics Correctness.* NeurIPS.
- Wang, X. et al. (2023). *Self-Consistency Improves Chain of Thought Reasoning in Language Models.* ICLR.
- Wang, Y. et al. (2024a). *Mixture-of-Agents Enhances Large Language Model Capabilities.*
- Wang, Z. et al. (2025). *Debate or Vote: Which Yields Better Decisions in Multi-Agent Large Language Models?* arXiv:2508.17536.
- Wu, Q. et al. (2024). *AutoGen: Enabling Next-Gen LLM Applications.* ICLR.
- Yule, G. U. (1900). *On the Association of Attributes in Statistics.* Phil. Trans. Royal Society A.
- Zheng, Y. et al. (2024). *HelloBench: Evaluating Long Text Generation Capabilities of Large Language Models.* arXiv:2409.16191.

Appendix A: Round Table Debate Data Flow

Figure 2 (Section 4.2) renders the single-round RAAC data flow: a task and the prior deliberation memory enter the Propose phase; proposals are cross-attacked in Challenge and answered in Defend (with the dropout check promoting a backup from a not-yet-represented provider); the Synthesizer integrates positions, attacks, and defenses into S^t; all models ratify; and the convergence test either emits S^t (updating durable deliberation memory) or opens round $t+1$, capped by the depth tier.

Appendix B: Deliberation Memory Schema

Persistent Deliberation stores structured JSON tracking unresolved tensions and per-model calibration across sessions:

```
{
  "version": "1.0",
  "session_id": "round-table-2026-02-16-001",
  "conclusions": [
    {
      "id": "C001",
      "proposition": "CDI must be measured across diverse domains.",
      "confidence": 0.95,
      "provenance": { "debate_round": 2, "ratified_by": ["m1", "m2", "m3"] },
      "status": "accepted"
    }
  ],
  "unresolved_tensions": [
    {
      "id": "T001",
      "description": "Bounding alpha theoretically vs. empirically.",
      "positions": {
        "m1": "Requires formal proof connecting to voting bounds.",
        "m2": "Empirical measurement suffices for evaluation."
      },
      "status": "open",
      "revisit_trigger": "Upon conclusion of empirical trials."
    }
  ],
  "model_calibration": {
    "Claude-Opus": {
      "strengths": ["formal_logic", "structural_analysis"],
      "reliability_by_domain": { "math": 0.85, "coding": 0.8 }
    }
  }
}
```