Council Intelligence: A Multi-Agent Architecture for Deep Reasoning

Whitepaper #2 (WP-002)

Dr. Narendra Gore

Atmakosh LLC

naren@atmakosh.com

December 6, 2025

Version 1.0.0

Abstract

Large language models (LLMs) are increasingly used as decision-support systems in domains where errors can be costly or harmful. Yet most deployed systems rely on a single model instance acting as an all-purpose cognitive engine. This monolithic pattern is misaligned with how deep reasoning unfolds in human practice, where insight typically emerges from interaction among multiple specialists, critics, and stakeholders. This whitepaper proposes *Council Intelligence*, a multi-agent architecture in which diverse internal agents—configured as distinct roles with explicit objectives and constraints—engage in structured debate, critique, and synthesis. Drawing on Indic epistemic traditions, especially Nyāya debate theory, we describe the conceptual foundations, architecture, evaluation metrics, and engineering blueprint for constructing council-based reasoning systems that are more robust, transparent, and culturally sensitive than single-agent alternatives.

Keywords: multi-agent systems, council intelligence, Nyāya, deep reasoning, AI safety, AtmakoshLLM

1. Introduction

Recent progress in large language models has transformed them from research curiosities into general-purpose reasoning engines embedded in workflows across law, medicine, finance, education, and governance. Despite this rapid adoption, there is an emerging consensus that today's systems often behave like talented but unreliable interns: fluent, fast, and creative, yet prone to hallucination, overconfidence, and shallow argumentation.

Most failures follow from a simple architectural fact: virtually all mainstream deployments rely on a *single* model instance to perform perception, retrieval, reasoning, evaluation, and ethical reflection in one continuous stream of tokens. The system is monological. It does not argue with itself, push back on its assumptions, or formally weigh divergent viewpoints before answering.

Human societies, by contrast, evolved practices that recognise the limits of single minds. Scientific peer review, judicial courts, parliaments, monastic debates, and philosophical disputations

all distribute reasoning across multiple agents. These agents hold different priors, incentives, and roles; the reliability of the outcome depends on the quality of interaction among them.

This whitepaper translates that social insight into a technical proposal. We introduce *Council Intelligence*: a multi-agent reasoning pattern where a "council" of specialist internal agents collaborate and compete to answer a query. Each agent is configured with its own role, strengths, and risk profile. A debate protocol allows them to challenge each other's claims and surface disagreements. A synthesis agent then integrates the resulting views into a final answer with explicit confidence and caveats.

The rest of the paper is organised as follows. Section 2 analyzes why single-agent LLM setups struggle with deep reasoning. Section 3 connects Council Intelligence to historical multi-agent reasoning in Indic traditions, especially the Nyāya school. Section 4 presents the architecture. Section 5 details internal agent roles. Section 6 describes the debate–consensus protocol. Section 7 introduces evaluation metrics. Section 7 walks through case studies. Section 8 outlines engineering implementation. Section 9 closes with future directions.

2. Why Single-Agent LLMs Fail at Deep Reasoning

LLMs are trained with autoregressive objectives that optimise local predictive accuracy: maximise the likelihood of the next token given the preceding context. While alignment finetuning and safety filters modify the surface behaviour, the underlying computational goal remains to continue text in a plausible way. This leads to several structural limitations.

2.1 Shallow optimisation over discourse

The model's loss function does not explicitly encode correctness of conclusions, faithfulness of citations, or adequacy of reasoning structure. If a fluent yet incorrect explanation appears plausible in the training distribution, the model has no intrinsic drive to reject it. The training objective rewards mimicry, not epistemic humility.

2.2 Lack of internal adversarial challenge

Human experts routinely test their ideas through conversation, debate, or internal dialogue. A single LLM instance executing a prompt has no counterpart that tries to refute its own arguments. Techniques like self-consistency or chain-of-thought help, but they are still different executions of the *same* policy, not distinct roles with different incentives.

2.3 No explicit modular roles

In real reasoning workflows, tasks such as retrieval, analysis, risk evaluation, and ethics review are often handled by different specialists. Single-agent LLM systems blur these roles. The same parameters that generate a scholarly explanation also decide how strongly to caveat medical advice. This can entangle capabilities and alignment pressures in brittle ways.

2.4 Limited perspective diversity

Deep understanding often requires combining multiple viewpoints: optimistic and pessimistic, technical and social, short-term and long-term. A single model, even if prompted to "consider

multiple perspectives," typically converges quickly to a single narrative thread.

These limitations motivate a shift from monolithic reasoning to multi-agent architectures that embed structured disagreement by design.

3. Multi-Agent Cognitive History: Nyāya and Beyond

Before computing, civilisations confronted the problem of reasoning reliability in social settings. Indic intellectual traditions developed rich frameworks that are surprisingly prescient for modern AI.

3.1 Nyāya debate as structured inference

The Nyāya school formalised rational debate with carefully defined roles and stages. A typical inference involved:

- **Pratijñā** (thesis): the claim under consideration.
- **Hetu** (reason): the supporting reason.
- **Udāharana** (example): an illustrative case.
- **Upanaya** (application): linking example to thesis.
- **Nigamana** (conclusion): the derived statement.

Debaters could challenge each step, accuse each other of specific fallacies, and appeal to shared *pramāṇas* (valid sources of knowledge). The aim was not merely to "win" but to expose weak assumptions and converge on defensible knowledge (Matilal, 1968).

3.2 Other Indic multi-agent analogues

Buddhist monasteries practised formalised debate where one monk challenged, the other defended, often under time pressure and with strict logical constraints. Jain philosophy advanced *anekāntavāda*, the doctrine of many-sidedness, arguing that complex objects can be legitimately described in partly incompatible ways depending on perspective. Mimāṃsa developed sophisticated hermeneutic rules for reconciling apparent contradictions in scripture through layered interpretation.

These systems share three meta-principles:

1. Reasoning is best done by multiple agents with different stances.
2. Protocols and shared rules matter as much as individual skill.
3. Truth can be multi-aspect; consensus should acknowledge residual uncertainty.

Council Intelligence is a computational realisation of these principles, tuned for modern LLM ecosystems.

4. Architecture of Council Intelligence

A Council Intelligence system comprises four main layers:

1. **Interface Layer** that receives user queries and returns final outputs.
2. **Evidence Layer** that retrieves, ranks, and organises external information.
3. **Agent Layer** composing multiple specialist LLM-based agents.
4. **Coordination Layer** implementing debate, scoring, and synthesis.

At a high level, the dataflow is:

User Query → Evidence Retrieval → Agent Deliberation (Debate & Critique) → Consensus Synthesis → Final Answer + Rationale

The system can be implemented as a microservice cluster or as an internal orchestration layer on top of a single LLM provider (Shoham and Leyton-Brown, 2009). Agents may share a base model but receive different system prompts, tools, and reward signals.

4.1 Evidence layer

The evidence layer typically combines vector retrieval over document embeddings with traditional symbolic lookups (SQL, knowledge graphs, APIs). Rather than giving every agent identical context, the coordinator can route different slices of evidence to agents depending on their role; for instance, a Safety Agent sees policy documents, while a Domain Expert sees technical manuals (Lewis et al., 2020).

4.2 Agent layer

Agents are stateless or stateful LLM wrappers with well-defined contracts:

- **Inputs:** query, evidence, transcripts of prior debate.
- **Outputs:** claims, justifications, numeric scores, and uncertainty estimates in a structured schema (e.g., JSON).

Agents differ along axes such as creativity vs. caution, local vs. global reasoning, or domain specialisation.

4.3 Coordination layer

The coordination layer mediates communication, enforces time budgets, scores arguments, and performs final synthesis. It can itself be implemented as a thin rules engine or as an additional LLM agent with a narrowly defined role.

5. Types of Internal Agents and Roles

While councils can be customised, a practical baseline includes the following roles.

5.1 Domain Expert Agent

Trained or prompted to focus on factual and domain-specific reasoning (e.g., cardiology, tax law). It seeks accurate, grounded explanations rather than speculative creativity.

5.2 Skeptic Agent

Tasked with finding flaws: missing edge cases, unstated assumptions, adversarial perspectives. It is explicitly rewarded for disagreement when justified by evidence.

5.3 Ethics & Impact Agent

Evaluates potential harms, fairness concerns, and cultural sensitivity. It reasons about stakeholders, norms, and long-term consequences, not just immediate correctness.

5.4 Pragmatist Agent

Focuses on feasibility, cost, operational risk, and implementation details. It may down-rank brilliant but impractical proposals.

5.5 Synthesiser Agent

Receives all previous arguments, along with confidence and critique metadata, and composes a coherent final answer. It must preserve nuance: acknowledging open questions and unresolved disagreement.

6. Debate, Consensus, and Synthesis Protocol

Council Intelligence is not just a set of agents; it is a *protocol* governing how they interact. A simple yet powerful pattern uses three rounds.

6.1 Round 1: Independent analysis

Each agent produces an initial analysis without seeing others' views. This reduces herd effects and preserves perspective diversity.

6.2 Round 2: Critique and counter-argument

Agents receive anonymised summaries of others' arguments and must:

- highlight specific points of agreement and disagreement,
- attach evidence or counter-evidence,
- rate each argument's strength and potential risk.

This round explicitly instantiates Nyāya-style *tarka* (critical reflection).

6.3 Round 3: Synthesis

The Synthesiser Agent aggregates:

- a ranked list of candidate conclusions,
- per-conclusion confidence scores,
- key evidence citations,
- flags from the Ethics Agent and Skeptic Agent.

The final answer delivered to the user may include a short "council transcript" exposing major disagreements, which improves transparency and trust.

7. Evaluation and Case Studies

To compare Council Intelligence systems empirically against single-agent baselines, we propose metrics such as Reasoning Depth Index, Evidence Grounding Score, Perspective Diversity Index, Cultural Alignment Score, and robustness under perturbation. These metrics can be estimated via human evaluation, automated checks, or hybrid pipelines, and applied to domains such as healthcare triage, regulatory impact analysis, and research hypothesis generation.

Early prototypes suggest that councils generate fewer but more robust hypotheses, identify edge-case risks more reliably, and produce explanations that domain experts find easier to audit compared to single-agent systems.

8. Engineering Implementation

A practical implementation pathway includes:

- exposing each agent as a stateless microservice,
- orchestrating them via a coordinator service,
- using declarative configuration for prompts and tools,
- enforcing a typed schema for claims and critiques,
- logging debates for auditability and learning.

This design supports horizontal scaling and heterogeneous backends, including mixing open-source and commercial models.

9. Future Directions

Promising directions include meta-learning of council composition, user-in-the-loop councils, cross-cultural ethical councils, and applications to long-horizon planning. Ultimately, reliable machine reasoning may resemble human institutions more than isolated savants. Council Intelligence offers a concrete, engineerable step in that direction.

References

Patrick Lewis, Ethan Perez, Aleksandra Piktus, et al. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *NeurIPS*.

Bimal Krishna Matilal. 1968. *The Navya-Nyāya Doctrine of Negation*. Harvard University Press.

Yoav Shoham and Kevin Leyton-Brown. 2009. *Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundations*. Cambridge University Press.