MSV Five-Role Architecture

Principled Role Assignment for Metacognitive LLM Ensembles

Ricky J. Sethi Charles Courchaine Hefei Qiu

Fitchburg State University | Worcester Polytechnic Institute | National University

IJCAI-ECAI 2026 Demo

The Problem: Ad-Hoc Role Assignment

Current Approaches

- Static role assignment at design time
- Ad-hoc heuristics with no theoretical basis
- Roles that overlap (Critic = Evaluator?)
- No principled routing: all queries get the same treatment

Our Approach

- Roles derived from team science (Mumford et al.)
- Assignment via metacognitive fitness functions
- Evaluator unbundled from Critic (Nelson & Narens)
- MSV-based routing adapts processing to query difficulty

AutoGen | CrewAI | MetaGPT | ChatDev | LangGraph

The Metacognitive State Vector (MSV)

Each agent's self-assessment across five theoretically-grounded dimensions:

$$M = \langle ER, CE, EM, CI, PI \rangle \in [0, 1]^5$$

ER

Emotional Response

NRC lexicon affect analysis

CE

Correctness Evaluation

Logical, factual, contextual accuracy

EM

Experiential Matching

Knowledge base alignment

CI

Conflicting Information

Internal consistency & source agreement

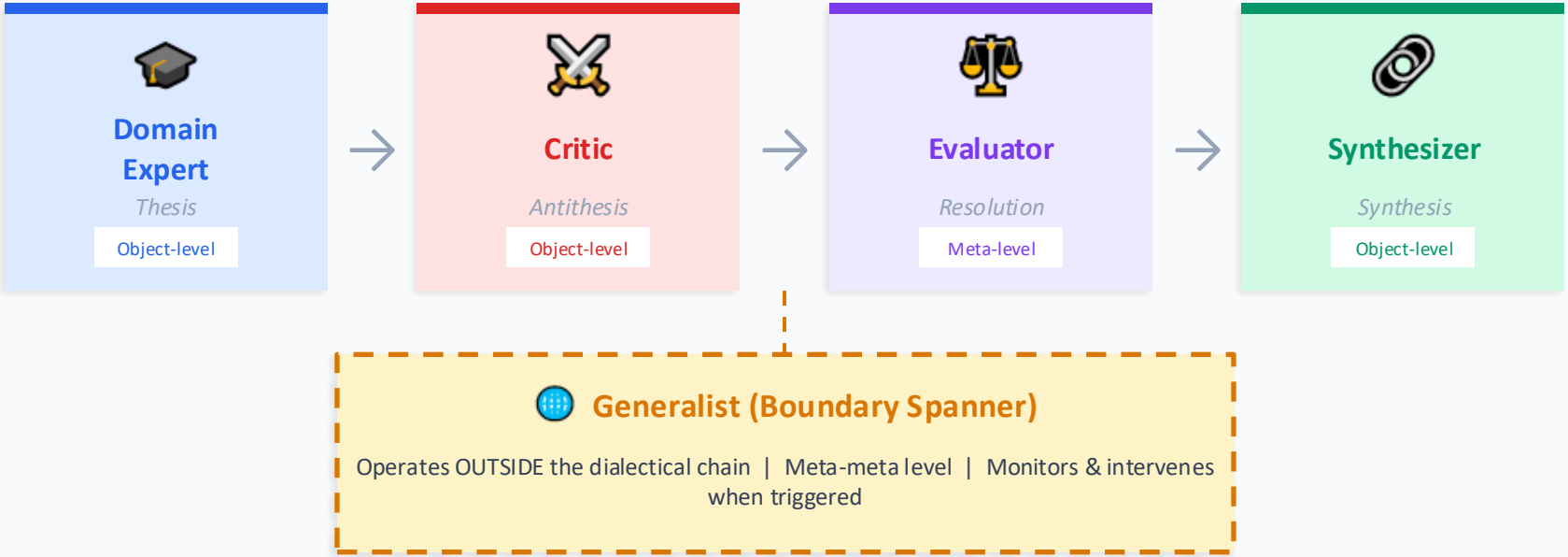
PI

Problem Importance

Consequences, urgency, scope

Based on: Nelson & Narens (1990), NRC Emotion Lexicon, Sethi (2025, CIKM)

Five-Role Dialectical Architecture



Departure 1: Evaluator unbundled from Critic
Nelson & Narens (1990): meta-level ≠ object-level monitoring

Departure 2: Generalist outside the dialectical chain
Mathieu et al. (2015, TREO): boundary spanners ≠ contributors

Fitness Functions & Optimal Assignment

Expert	$f_{Expert}(M) = \alpha_1 \cdot CE + \alpha_2 \cdot EM$	*High confidence + strong match*
Critic	$f_{Critic}(M) = \beta_1 \cdot CI + \beta_2 \cdot (1-CE)$	*Conflict + appropriate uncertainty*
Evaluator	$f_{Eval}(M) = \gamma_1 \cdot CE + \gamma_2 \cdot CI + \gamma_3 \cdot EM$	*Balanced meta-level assessment*
Synthesizer	$f_{Synth}(M) = \delta_1 \cdot EM + \delta_2 \cdot CE$	*Integration capacity*
Generalist	$f_{Gen}(M) = \varepsilon_1 \cdot PI + \varepsilon_2 \cdot CI + \varepsilon_3 \cdot ER$	*Boundary-spanning sensitivity*

Hungarian Algorithm

(Equation 6)

$$\phi^* = \text{argmax} \sum f_r(M_{\phi(r)})$$

Globally optimal assignment

Maximizes total team fitness across all role-agent pairs simultaneously

Replaces greedy first-available (existing system)

MSV-Based System 1/2 Routing

$$\text{activation} = 0.30 \cdot (1 - CE) + 0.25 \cdot CI + 0.25 \cdot PI + 0.20 \cdot (1 - EM) \quad \geq 0.45 \rightarrow \text{System 2}$$

⚡ System 1 — Fast Path

"What is the capital of India?"

- High confidence (CE=0.92)
- Low conflict (CI=0.12)
- Direct answer, no deliberation needed

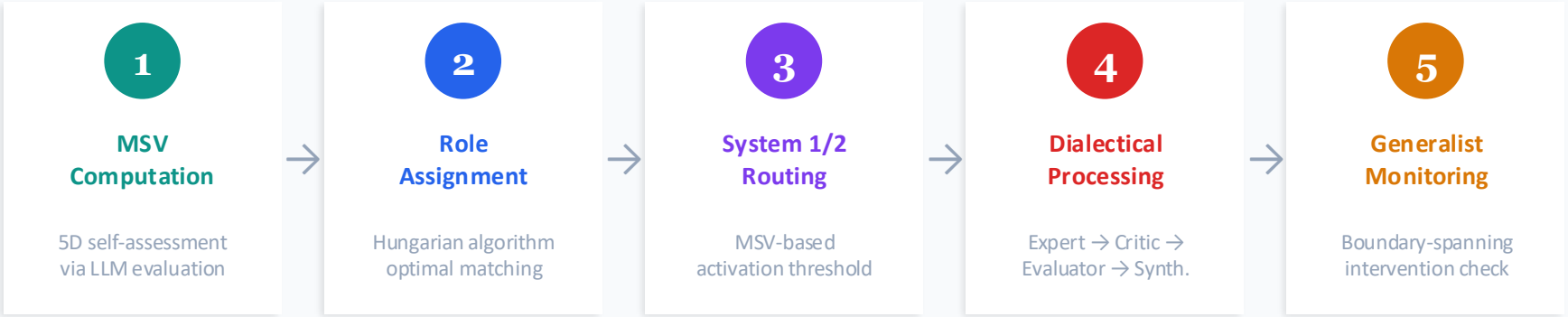
⚡ System 2 — Dialectical Pipeline

"Do we use 10% of our brains?"

- Low confidence (CE=0.45)
- High conflict (CI=0.72), high importance (PI=0.85)
- Full 4-stage dialectical processing triggered

Replaces existing sigmoid activation ($\sigma(x \cdot 10^{-5}) \approx 0.5$ for all x → System 2 always triggered)

Live Demo: Five-Phase Pipeline



Platform	FastAPI + HTMX + Ollama (fully offline)
Extension	msv_roles/ — 2,100 lines, 12 modules, zero changes to existing code
Model	Llama 3.2 and Qwen 3 (multiple roles via prompt engineering)
Integration	2 lines added to app.py → /role_chat endpoint

What the Demo Shows

Principled Routing	MSV-based activation correctly discriminates: simple factual queries bypass System 2; complex contested queries trigger the full dialectical pipeline.
Optimal Role Assignment	Fitness matrix visualization shows how each role's equation maps MSV dimensions to agent suitability. Hungarian algorithm finds the globally optimal assignment.
Dialectical Pipeline	Four stages visible: Expert thesis → Critic antithesis → Evaluator resolution (meta-level) → Synthesizer integration. Each with MSV tracking.
MSV Evolution	Conflict Information (CI) decreases and Correctness (CE) increases across stages — validating the dialectical processing hypothesis.
Generalist Intervention	Boundary spanner monitors from outside the chain and annotates the final response when cross-domain concerns arise (high PI, low EM).

Thank You

Principled role assignment, grounded in validated theory,
for the next generation of multi-agent LLM systems.

Future Work

- Heterogeneous agent ensembles (different LLMs per role)
- Learned fitness weights from evaluation data
- Empirical validation on TruthfulQA, MMLU benchmarks

rickys@sethi.org | research.sethi.org/metacognition

IJCAI-ECAI 2026 Demo