METACOGNITIVE AGENTS · CONFLICT FORECASTING

Interventionist Metacognition

Implementing Nelson–Narens monitoring–control theory in an LLM agent

Poli Nemkova, PhD · University of North Texas · Lab42 Research Group · poli.nemkova@unt.edu

Nelson–Narens theory | uncertainty taxonomy | interventionist self-regulation | LLM agent architecture

01 The forecasting problem

Predicting armed-conflict escalation means reasoning under **severe epistemic uncertainty** — incomplete evidence, contradictory sources, and structural novelty that standard LLM inference handles poorly.

Our response is a **metacognitive agent** that intercepts its own inference **before committing a prediction** — classifying the epistemic state of each input and routing execution accordingly.

02 Two levels, two flows

META LEVEL
Models and regulates the object level

monitor · object→meta ↑ ↓ control · meta→object

OBJECT LEVEL
Generates the prediction

Flows are **asymmetric & sequential**: monitoring conditions control; control intervenes on execution rather than revising output. We extend this with Minsky's *administrative agency*, operationalized intra-agent.

03 A six-stage cognitive loop

Object-level execution is **gated by the metacognitive stages** — pre-generative gating that is absent from chain-of-thought and self-reflection architectures, which conflate object- and meta-level reasoning in a single undifferentiated pass.

Metacognitive · meta-level Object-level · gated execution

GATE

STAGE 1	STAGE 2	STAGE 3	STAGE 4	STAGE 5	STAGE 6
FORECAST	**MONITOR**	**CONTROL**	**EXECUTE**	**SYNTHESIZE**	**GATHER**
Draft an initial prediction	Classify into 7 typed states	Type-conditioned routing directive	Run the routed recovery procedure	Commit a calibrated prediction	Collect evidence for next cycle

← GATHER feeds the next FORECAST — a closed self-regulatory loop

04 Typed uncertainty → recovery routing

UNCERTAINTY STATE	ROUTING DIRECTIVE
INSUFFICIENT_EVIDENCE	→ Targeted evidence gathering
CONFLICTING_SOURCES	→ Evidence triangulation
DISTRIBUTION_SHIFT	→ Re-anchor to current regime
LOW_QUALITY_DATA	→ Source re-weighting
TEMPORAL_INCONSISTENCY	→ Temporal re-anchoring
MODEL_AMBIGUITY	→ Principled abstention
CONFIDENT	→ Proceed to execution

05 Results

0.91 F1 · TRACTABLE CASES — roughly half the set; principled abstention on the rest.

775 prediction cases

5 conflict contexts

CALIBRATION — BRIER SCORE
Lower is better

Metacognitive agent (OURS)	0.393
Few-shot LLM (BASELINE)	0.473

06 Ablation — what drives the gain

Remove routing type-specificity while preserving the monitoring stage, and calibration gains **collapse to baseline**. Monitoring without typed control is insufficient — **administrative redirection of reasoning**, not richer generation, drives the improvement.

CONTRIBUTION

First architecture to implement **both Nelson–Narens flows mechanistically** — monitoring and typed control as programmatic operations — gating generation **before it commits**.