A Metacognitive Upstream Routing Framework for Accuracy Preservation and Computational Efficiency in Artificial Intelligence Systems

Naavya Shetty
University of Illinois Urbana-Champaign
shetty.naavyasukesh@gmail.com

Naavya Shetty

shetty.naavyasukesh@gmail.com

University of Illinois Urbana-Champaign

- B.S. Computer Science & Philosophy
- Distinguished Undergraduate Researcher awardee 2025-2026

Research Interests

- Cross-disciplinary – Philosophy & Cognitive Psychology in AI
- e.g. decision-making under uncertainty; resource-aware AI systems; interpretability and evaluation; human-AI alignment

Agenda

1

The Problem

Current AI systems compute too much and too indiscriminately

2

Our Proposal

PMS 2.0 – Architecture and Routing Logic

3

Results & Trade-offs

Numbers, conclusions, and everything in between

4

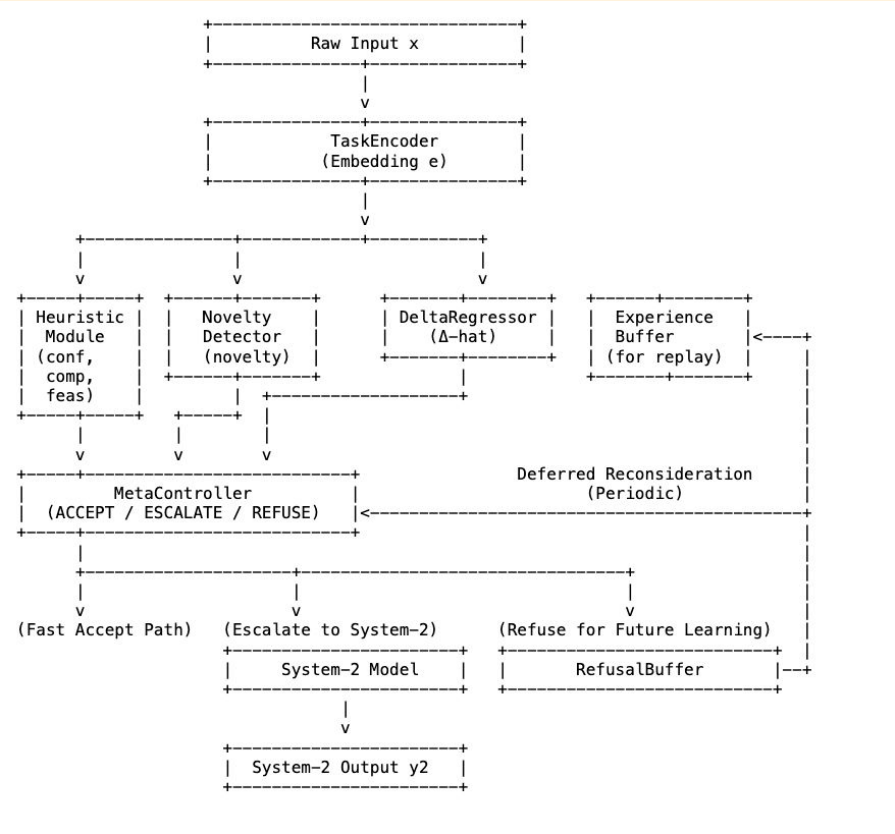
Future Plans

Action plans to address current limitations

The Problem

- Input arrives → model engages → output produced
- No interrogation, consideration, or a *no*
- Predictable failure modes:
  - wasted compute on trivial inputs;
  - false confidence on out-of-distribution ones;
  - no engagement trace on malformed inputs.
- Missing gating mechanism – upstream, assess, justify
- Operationalising fast appraisal *before* slow deliberation

Our Proposal



Results & Trade-offs

- 200 held-out synthetic inputs; controlled setting
- Baselines: unconditional System-2 invocation, threshold-based gate, heuristic-only variant
  - **~86% reduction** in downstream System-2 invocations (120 → 17)
  - **Conditional accuracy preserved** on escalated inputs
  - **Fully auditable decision logs** without modifying System-2
- Aggressive savings reduce coverage; PMS 2.0 makes this explicit and measurable.

Future Plans

- Adaptive threshold calibration & learned MetaController
- Refined benefit estimation
- Multi-tier routing
- Fully online learning