AEMA: Verifiable Evaluation Framework for Trustworthy and Controlled Agentic LLM Systems

Yen-Ting Lee, Keerthi Koneru, Zahra Moslemi, ***Sheethal Kumar, Ramesh Radhakrishnan***

Evaluation Should act as a control mechanism not just a score

Motivation

- Agentic LLM systems reason, plan, and act over **multiple steps**
- Traditional evaluation focuses on **single responses**

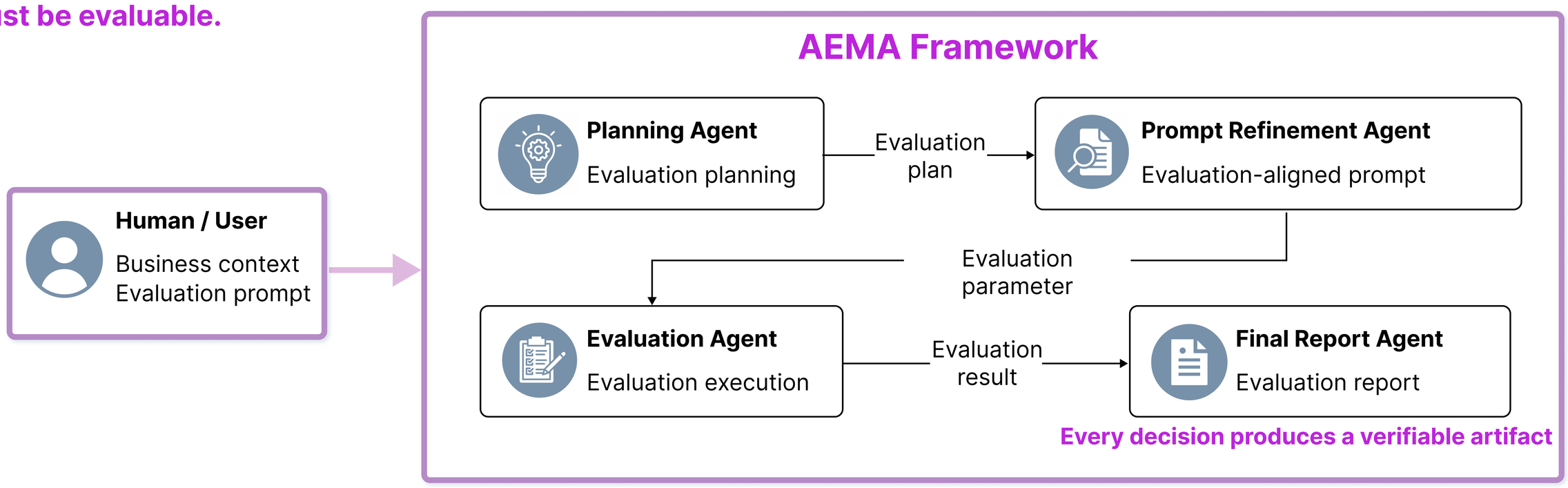
Why This Matters:

- Enterprise and high-stakes workflows require:
  - **Repeatability**
  - **Transparency**
  - **Verifiable decision traces**

! Outcome accuracy alone is insufficient: **the process must be evaluable.**

Enterprise Trust Angle

- Evaluation becomes **a governance and control layer**
- Enables:
  - Fixed thresholds
  - Reduced manual overrides
  - Regulatory auditability
- Moves agentic AI closer to **trustworthy deployment**



Key Idea: What is AEMA

- **AEMA (Adaptive Evaluation Multi-Agent)**
  - **Process-aware**, **multi-agent evaluation** framework
  - Plans, executes, and aggregates **step-level evaluations**
  - Produces **auditable**, **reproducible** evaluation traces
  - Operates under **human oversight**

! From single-shot scoring to multi-step verifiable evaluation workflows.

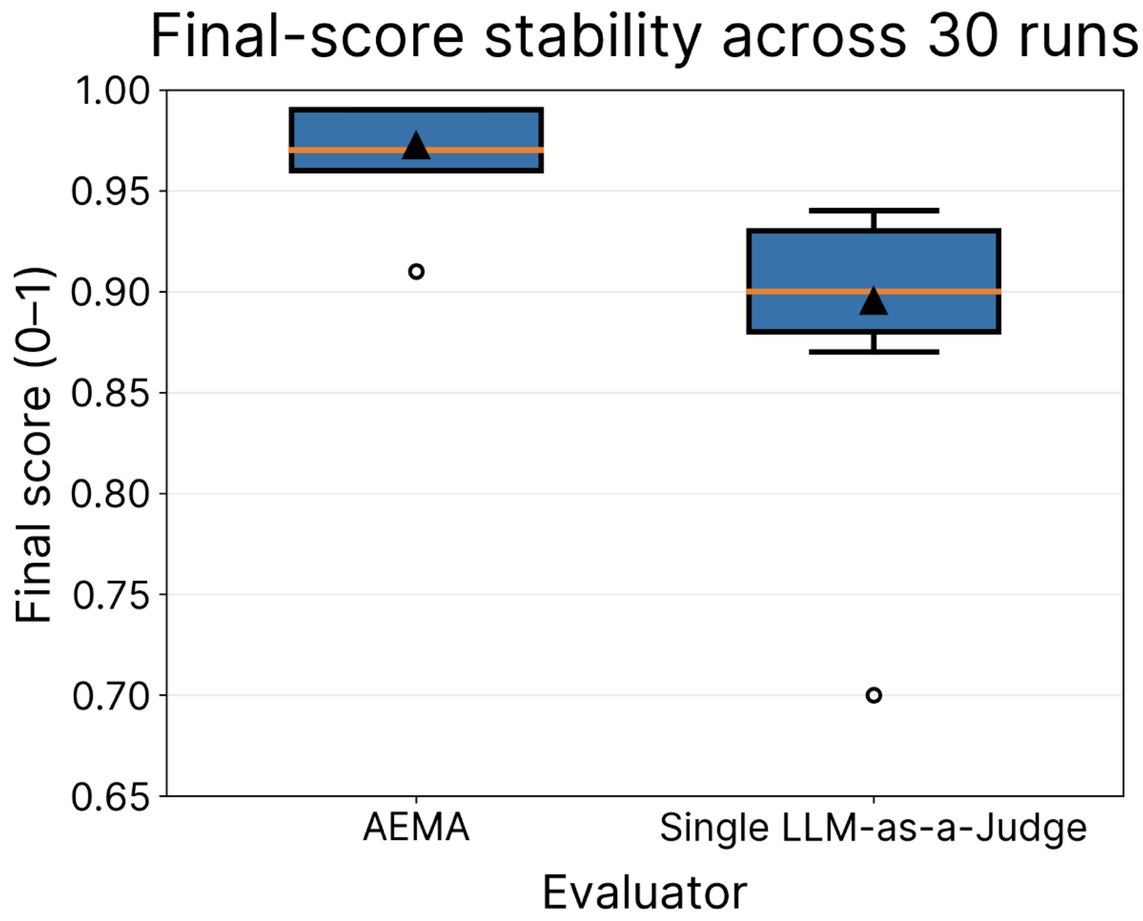
Key Findings

- **Lower score dispersion**
  - AEMA produces **stable evaluations** across runs
- **Stronger human alignment**
  - Smaller absolute error vs. human judgment
- **Robustness to noise**
  - Maintains alignment even on degraded inputs

! **AEMA delivers more stable, human-aligned, and auditable evaluations than a single LLM-as-a-Judge.**

Why AEMA is Different?

Single LLMJ	AEMA
• Single response score • High Variation • Opaque Reasoning • No adaptation • Hard to oversee	• Step-level + end-to-end evaluation • Low dispersion, stable scores • Auditable evaluation logs • Domain and task-aware planning • Human-in-the-loop by design



Key Contributions

1. **Process-aware evaluation framework** for agentic LLM systems
2. **Multi-agent evaluators** with adaptive planning and debate
3. **Verifiable evaluation traces** supporting human oversight
4. **Empirical evidence** of stability and alignment gains