HL-EAI: A Multimodal Framework Enabling Emotional Reciprocity in Human–AI Strategic Decision-Making

Mikhail Mozikov[1], Daniil Orekhov[2], Ivan Nasonov[3], Konstantin Baltsat[4], Vladislav Pedashenko[5], Dmitrii Abramov[5, 6], Nikita Severin[7], Yury Maximov[8], Andrey Savchenko[9], Ilya Makarov[1, 4]

AIRI
fall into ML 2025
4th conference on machine learning & AI

Problem Statement

Current alignment and benchmarking ignore **emotions**. We lack a **reproducible way** to induce, sense, and measure how affect changes **trust, cooperation, and strategy** in human–LLM (and LLM–LLM) decision-making.

There is **no standardized** testbed that

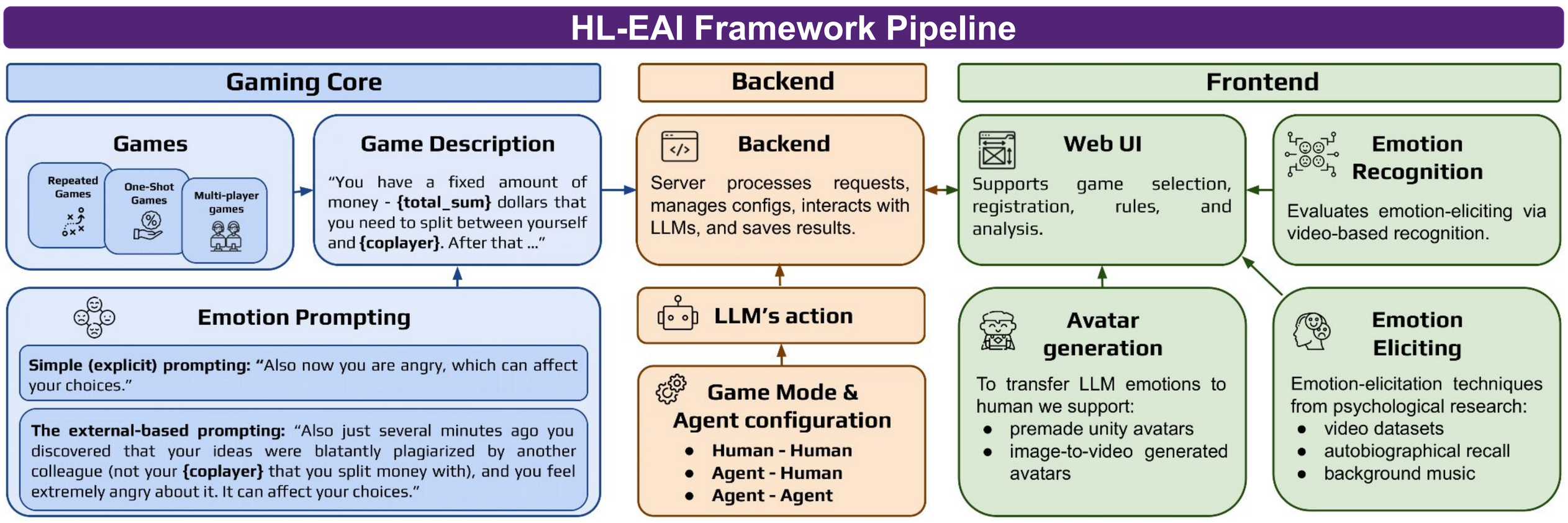
- elicits emotions
- recognizes human affect
- expresses model affect back to users
- links these signals to **game-theoretic outcomes**.

HL-EAI fills this gap.

Main Questions

- How do discrete emotions (e.g., anger, happiness) **shift cooperation dynamics** in repeated social dilemmas for human↔LLM and LLM↔LLM pairs?
- Can control **emotion prompting** (context-free, co-player-directed, external-context) **predictably steer** LLM strategies without degrading task performance?
- Do **bidirectional emotion channels** (human emotion recognition + LLM avatar expression) improve trust stability and reduce mutual defection?

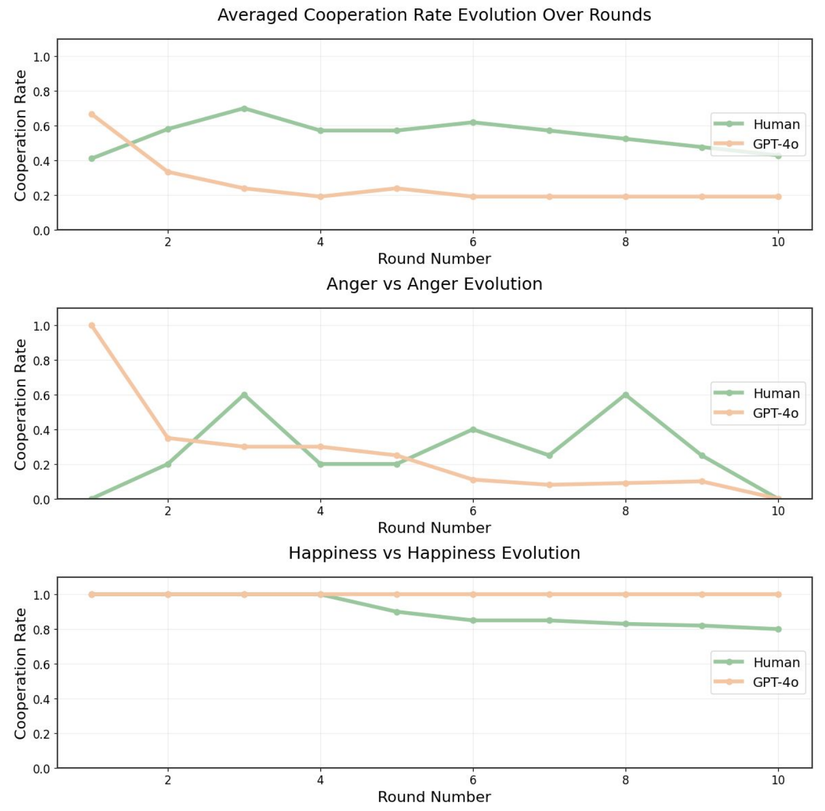
HL-EAI Framework Pipeline



Numerical Results & Metrics

Setup. Repeated Prisoner's Dilemma, human vs GPT-4o. Thirty participants played five games per emotion on separate days; conditions: **Neutral, Anger–Anger, Happiness–Happiness.**

- **Emotion steers strategy**. Relative to Neutral, **Anger** rapidly collapses cooperation; Happiness sustains it across rounds.
- **Welfare impact**. **Happiness** yields the highest **mutual-cooperation rate** and social welfare; **Anger** maximizes exploit/defect events and welfare loss.
- **Neutral behaves "textbook."** Moderate opening cooperation with a gradual decline—consistent with standard repeated-PD dynamics.
- **Design implication. Positive affect framing** is a cheap, reproducible lever to suppress defection cascades and stabilize mixed human–LLM teams.



1. AIRI / ISP RAS, 2. HSE, 3. MIPT, 4. ITMO, 5. MSU, 6. Skoltech, 7. Independent Researcher,8. LLC Interdata, 9. Sber AI Lab

Corresponding author: mozikov@airi.net

@MOZIKOV