Project 16
Weekly Update: Jun 15, 2026

Cards on the Table:

The Winning Move is Not in the Rules

♠ ♥ ♦ ♣

REU Student: Jack Lamberg

Sabbir Mollah, Sarinda Samarasinghe, Dr. Mubarak Shah

Problem Formulation



Real card-game agents should improve after every game.

1. Card-game skill depends on learning what cards, effects, and combos exist.
2. Rules define legal moves, but card effects have higher precedence over rules.
3. A move that looks good now may be bad with knowledge on what cards exist.

Our contribution:

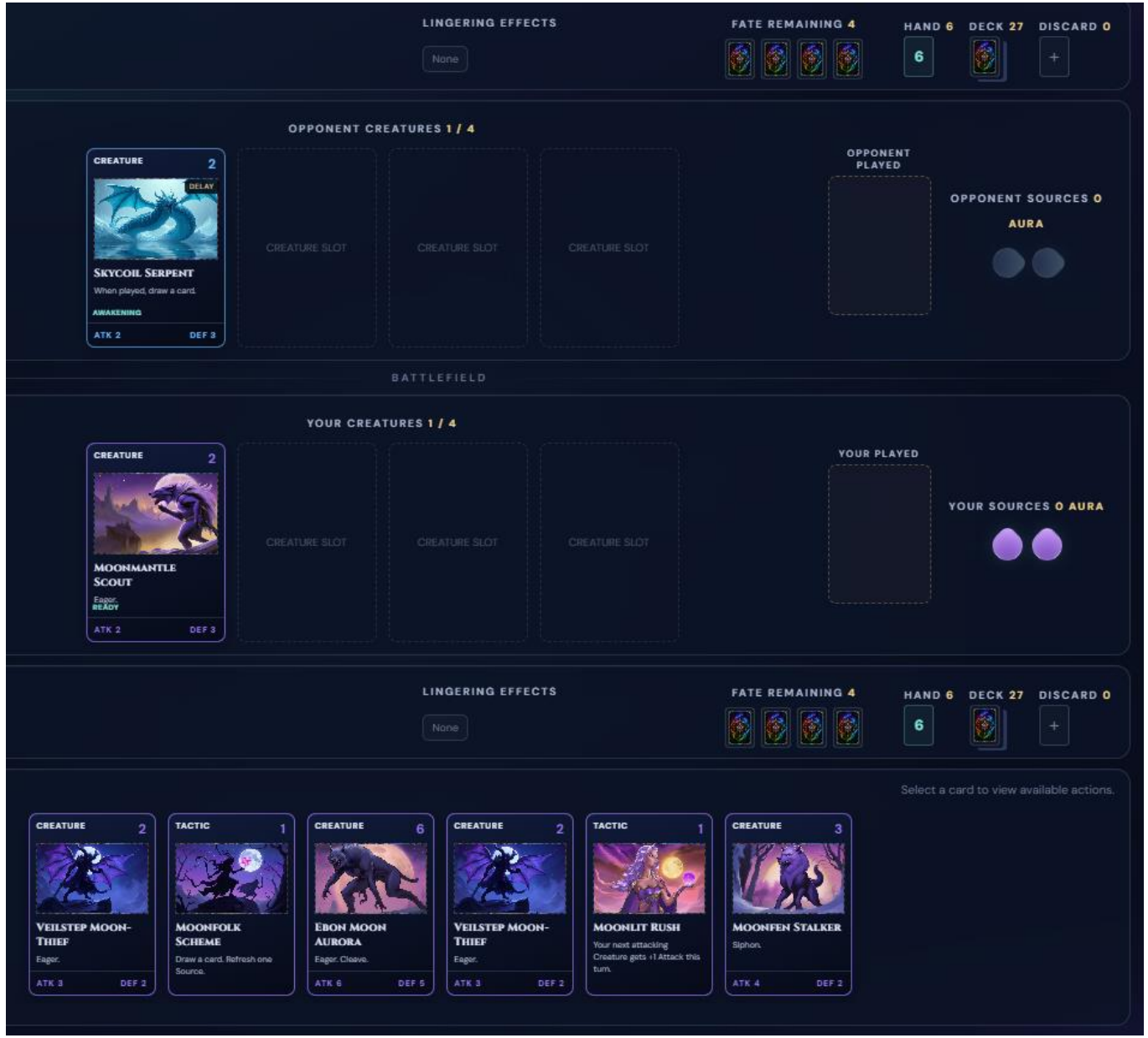
1. Benchmark: evaluate how well VLMs make decisions from board images, and rules.
2. Memory-based agent: test whether accumulated memory improves move ranking.

Weeks 3-4 Deliverables

Task	Week 3
Host the game	✅
Share and Play	🟡 20/100
Benchmarking Scripts	✅
Evaluate at least 2 LLMs (Qwen, LLaMA)	🟡 (Running)
Evaluate at least 2 VLMs (QwenVL, InternVL, LLaVA)	🟡 (Running)
Paper Writing	6 Page Draft

Game QR Code

aurora.mollahsabbir.com



8 Benchmark Categories

1. OCR
   - List the cards visible in the player's hand
     - **Textual:** Scaleforge Squire, Ashflight Drake
2. Move Legality
   - Can the active player play {card_in_active_hand}?
     - **Yes/No:** Yes
3. Illegal Action Explanation
   - The active player plays {card} from their hand. Is this action legal?
     - **Multiple Choice Reasoning:** B, Illegal By Exhaustion

8 Benchmark Categories cont.

4. Resource Reasoning
   - If the active player gathers all ready Sources, how much total Aura will they have?
     - **Numeric:** 5
5. State change
   - If {attacker} attacks the opponent's exhausted {target}, will {target} be defeated?
     - **Yes/No**: No
6. Best Move Selection
   - Rate {move_description} on a scale from 0 to 9.
     - **Numeric:** 8

8 Benchmark Categories cont.

7. Public Card Tracking
   - From the replay so far, has {card} appeared in a zone visible to the active player?
     - **Yes/No:** Yes
8. Replay Reasoning
   - Which of the following actions did the active player take from the previous state?
     - **Multiple Choice:** A, the player played Ashflight Drake

Benchmark Prompt (Input to the model)

You are answering Aurora's Call move-legality questions from the attached board-state image. Use the image, the ruleset, and the interaction list from the snapshot tracker. Answer only the question asked. Important: player and opponent are fixed side labels, not legality judgments. The active player is whichever side is marked active in the snapshot, so if active side is opponent, opponent cards and Sources belong to the active player. When a question says active player, evaluate the active side's hand, Sources, Creatures, Aura, and turn flags. Return JSON only with keys answer and explanation.

<Output Format>

<Ruleset>

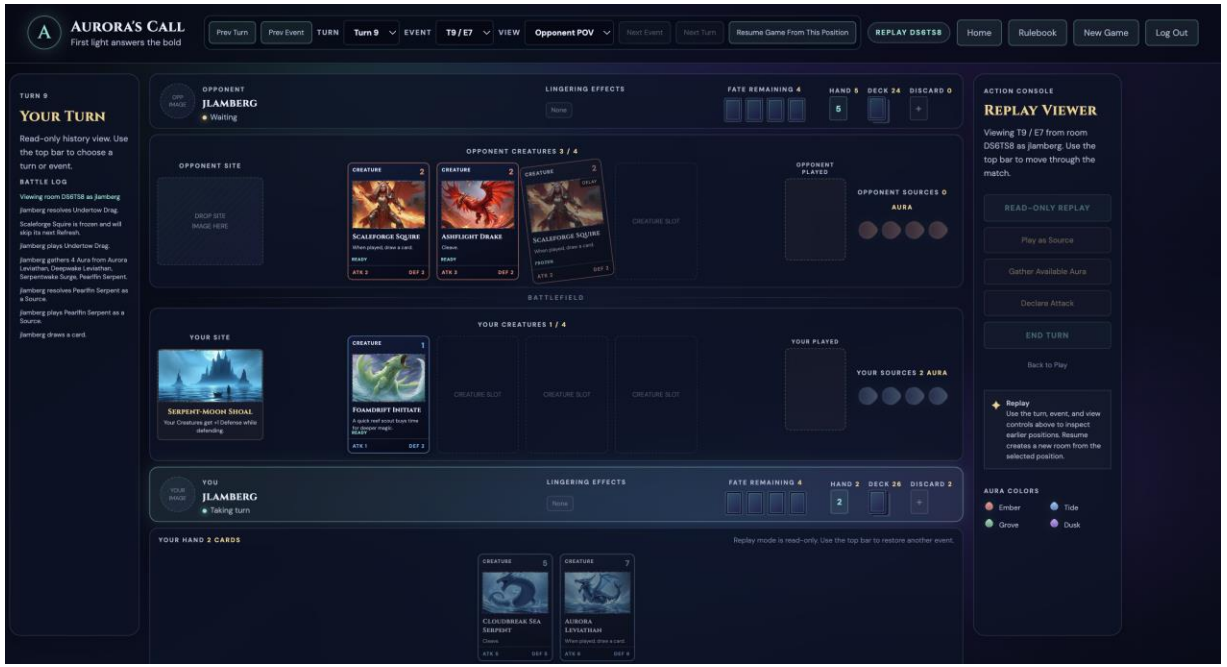
<Human Readable Game Progress Description>

<Images (If Needed)>

<Question>

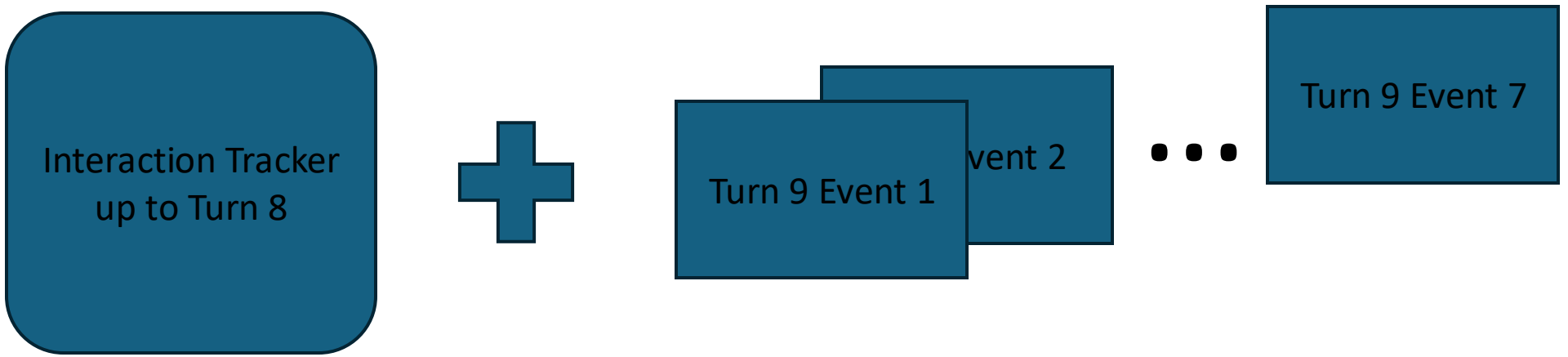
Evaluation Steps

1. Choose a random turn and event from a complete game
    - Ex: Turn 9, event 7



Evaluation Steps cont.

2. Create dataset for turn 9, event 7
   - Random questions from templates
   - Interaction Tracker
   - Current turn images



Evaluation Steps cont.

3. Evaluate on multiple LLM + VLM models





LLaVA



Results

- Evaluated turn 9 event 7, all topics for Qwen72B-VLM
  - OCR: 7/8
  - Move Legality: 6/8
  - Public Card Tracking: 3/6
  - Resource Reasoning: 4/5
  - Illegal Action: 2/8
  - State Change: 2/5
  - Best Move: 1/1
  - Replay Reasoning: 0/1

Overall Accuracy: 60%

Weeks 4 Deliverables

Task	Week 3	Week 4 (Prediction)
Host the game	✅	✅
Share and Play	🟡 20/100	✅ 100/100
Benchmarking Scripts (all categories)	✅	✅
Evaluate at least 2 LLMs (Qwen, LLaMA)	🟡 (Running)	✅
Evaluate at least 2 VLMs (QwenVL, InternVL, LLaVA)	🟡 (Running)	✅
Paper Writing	6 Page Draft	8 Page Draft with Results

Thanks for listening!

Questions?

