Project 16
Weekly Update: Jun 8, 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:

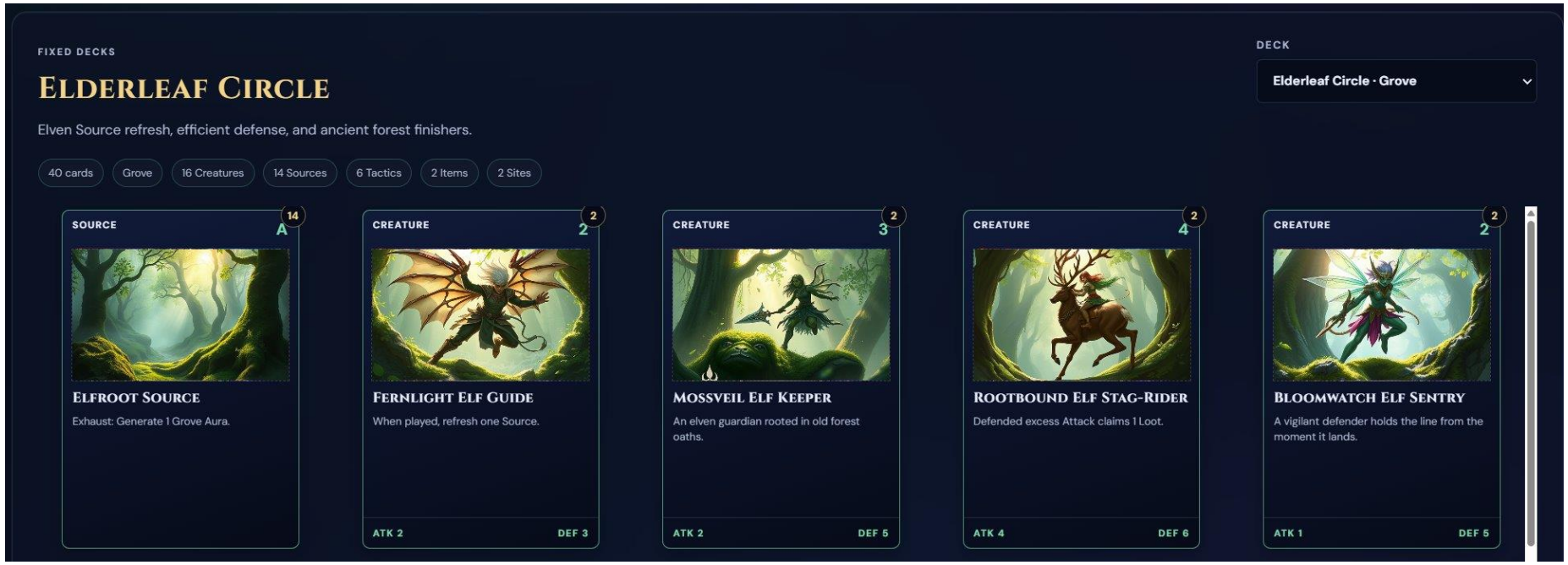
1. Benchmark: evaluate how well VLMs make decisions from board images, and rules.
2. Memory-based agent: test whether accumulated memory improves move ranking.

Weeks 1-2 Deliverables

Task	Last Week	This Week
Create Ruleset	✅	✅
Create Game Engine	✅	✅
Create initial balanced card decks + Generate artwork	🟡	✅
Implement dataset schema	✅	✅
Collect initial toy dataset	✅	✅
Finalize game based on initial gameplay	🟡	✅
Game data -> Benchmark data generation logic	🟡	✅

Card Generation

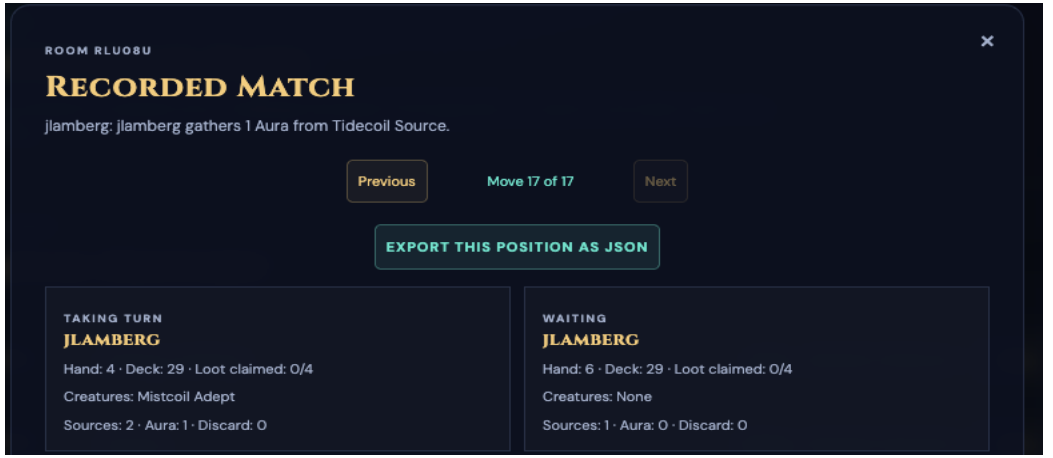
1. 160 Cards, 4 Decks
2. Generated artwork using Stable Diffusion XL

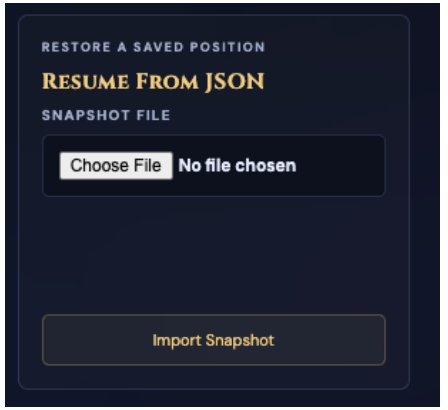


JSON Game States Dataset

Reproduce games using the JSON

- Export a game's JSON from any place in the game replay
- Import a game JSON to return to the current state
- Allows playing a game from the middle to create new states





8 Benchmark Categories

1. OCR - VLM
   - Identifying the current board state
2. Move Legality - VLM + LLM
   - Identifying if a move is legal/illegal
3. Illegal Action Explanation – VLM + LLM
   - Reason why a proposed action is illegal

8 Benchmark Categories cont.

4. Resource Reasoning - VLM + LLM
    - Reasoning for energy-based moves
5. State change – VLM + LLM
    - Explaining how playing a card affects the game
6. Best Move Selection – VLM + LLM
    - Rating random moves and actual moves

8 Benchmark Categories cont.

7. Public Card Tracking - VLM + LLM
    - Evaluating card presence recall
8. Replay Reasoning – VLM + LLM
    - Explain how to get from state A to state B

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 State Description>

<Image (If Needed)>

<Question>

Benchmark Prompt cont.

- **<Output Format>** The answer must be exactly yes or no. The explanation must be one short sentence. Example: {"answer":"yes","explanation":"The active player has enough Aura and the action is allowed now."}
- **<Human Readable Game State Description>**
    - 15: sarinda (opponent) plays Moonwell Source.
    - 16: sarinda (opponent) gathers 2 Aura.
    - 17: sarinda (opponent) ends the turn.
    - 18: jlamberg (player) draws a card.
        - hand_card jlamberg (player): name=Measured Undertow | type=tactic | faction=Tide | cost=2 | text=Your Creatures get +1 Attack this turn.

Benchmark Prompt cont.

- **<Question>**
  - **Category: Move Legality** - Can the active player play Moonshade Archivist?
    - Ground-truth: Yes
  - **Category: OCR** – What creatures are in the active player's hand?
    - Ground-truth: Gloamwing Moon Duelist, Moonmantle Scout, Starveil Mooncloak, Moonshade Archivist, Moonwell Source
  - **Category: Public Card Tracking** – Have you seen the Moonmantle Scout card in this game?
    - Ground-truth: No

Results

- Move Legality Prompts
  - Model: QwenVL 72B
    - Input: ruleset, interaction tracker, game state image, question
    - Accuracy: 57% for 1 state, 4/7 questions answered correctly
  - Model: QwenLLM 72B
    - Input: ruleset, interaction tracker, question
    - Accuracy: 64% over an entire game

Week 3-4 Plan

Task	This Week
Host the game	TODO
Share and Play	TODO
Convert to Benchmark data	TODO
Evaluate at least 2 LLMs	TODO
Evaluate at least 2 VLMs	TODO

Thanks for listening!

Questions?

