Project 16
Weekly Update: Jun 25, 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:

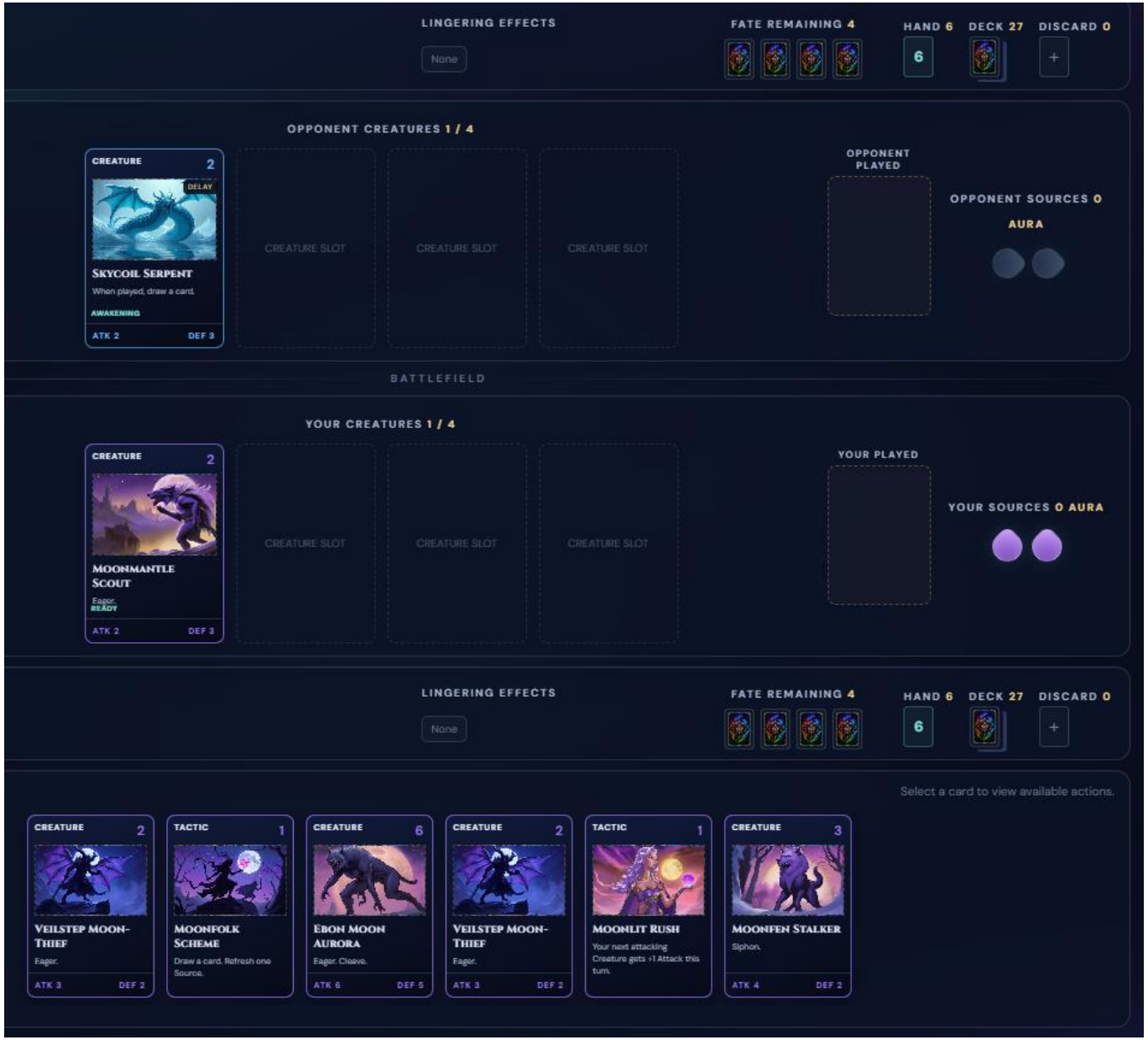
1. LLM/VLM Understanding: Evaluate LLM/VLM performance on in-game questions.
2. LLM/VLM Gameplay: Evaluate LLM/VLM performance in playing full games.
3. Memory-based agent: Evaluate impact of accumulated memory on these metrics.

Weeks 4-5 Deliverables

Task	Week 5
Host the game	✅
Benchmarking Scripts	✅
Share and Play	🟡 20/100
Develop 4 strong bots to effectively play the game	✅
Generate a game dataset using bot games	✅ (Maybe want more data)
Evaluate Model Understanding on 2 LLMs & 3 VLMs	✅ (Qwen, LLaVA/Llama, Intern…)
Evaluate Model Gameplay on 2 LLMs & 3 VLMs	✅
Literature Review on Memory	🟡

Game QR Code

aurora.mollahsabbir.com



Understanding: 7 Benchmark Categories

1. **OCR** – List the cards in the player's hand.
2. **Move Legality** – Can the player play {card_name}?
3. **Illegal Action Explanation** – Can the player play {card_name} as a source?
4. **Resource Reasoning** – Can the player afford {card_name}?
5. **State change** – Who would survive if {card_name} attacked {card_name}?
6. **Public Card Tracking** – Has {card_name} been seen in the game so far?
7. **Replay Reasoning** – From the previous state, what action was taken by the player?

* Removed "Best Move" topic as this evaluation is done when the models play the game.

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>

Game Understanding Evaluation Results

	OCR	ML	IA	PCT	RR	AR	ST	Overall
QwenLLM-32B	60.51%	**58.39%**	47.83%	50.33%	**79.79%**	33.21%	44.55%	50.92%
Llama-8B	51.65%	51.24%	42.36%	47.00%	62.77%	32.29%	31.72%	44.42%
LLaVA-7B	53.62%	49.00%	32.61%	41.21%	15.29%	27.69%	27.88%	39.27%
InternVL-38B	**83.33%**	39.92%	**63.25%**	**63.67%**	54.26%	29.91%	67.07%	**57.53%**
QwenVLM-32B	75.53%	42.00%	52.17%	57.83%	69.15%	**38.35%**	**69.25%**	55.39%

- Shows accuracy (% correct) of each model per topic and overall
- Notes:
  - Bigger models perform better
  - LLM and VLM are better at different tasks

Gameplay Bots

- Evaluation dataset no longer requires humans to play
- LLM/VLMs being evaluated will play against the same bots

- Smart bots:
    1. **Random** – Sample from legal set of moves
    2. **Rule-based** – Predefined sequence of actions
    3. **Greedy** – Move based on One-step lookahead score
    4. **Monte Carlo Tree Search (MCTS)** – Move based on 150 random step lookahead

Model vs Bot (48 Games per matchup)

Win Percentage					
	Random	**Rule-Based**	**Greedy**	**MCTS**	**Overall***
QwenLLM-32B	83.3%	10.4%	16.7%	(Running)	36.80%
Llama-8B	29.2%	4.2%	8.3%	(Running)	13.90%

- We play 48 games for each model/bot pairing
- Random bot loses most often as expected
- LLaMa-8B has a lot of trouble with this task

Weeks 6-7 Deliverables

Task	Week 6-7
Share and Play	(20/100 games)
Generate larger dataset for game understanding questions	
Complete evaluation on larger dataset	
Evaluate model vs MCTS	(Running)
Play language models against each other	
Literature review on memory	
Implement memory module using collected data	

Thanks for listening!

Questions?

