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.

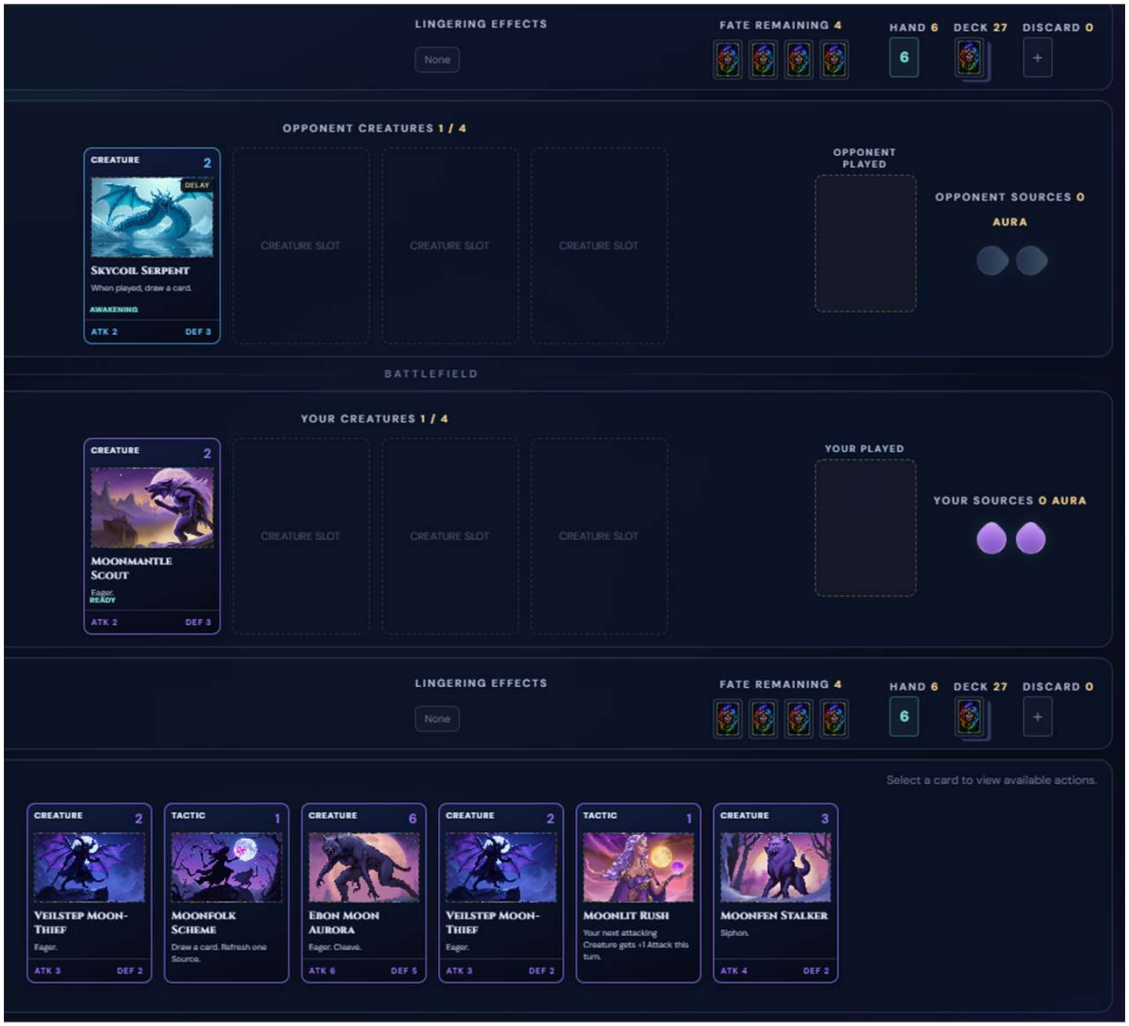
1. Card-game skill depends on learning what cards, effects, and combos exist.
2. Rules define legal moves, but card effects have higher precedence over rules.
3. A move that looks good now may be bad with knowledge on what cards exist.

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.

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?

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

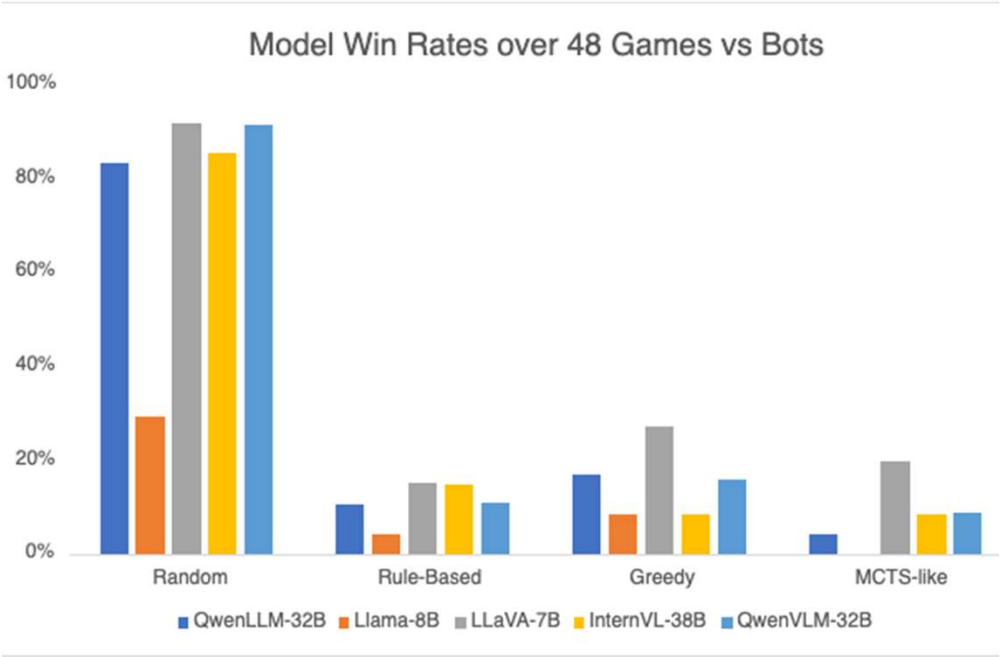
Understanding Evaluation

	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%

- Ask per topic questions from game states
- Shows accuracy (% correct) of each model per topic and overall
- LLMs & VLMs are better at answering different questions

Gameplay Evaluation

- We play 48 games for each model/bot pairing
- Models struggle against any hint of a strategy
- LLM & VLM performance is similar



Our Memory Perspective- Three Stages

- **Build stage**:
  - Individually build memory documents from games
- **Combine stage:**
  - ADD/MERGE/UPDATE documents sequentially
- **Retrieval stage:**
  - Use board state tag to retrieve relevant memory to the current game state

- Need expert level games to use for memory build stage
- Evaluate expert level memory against all four bots

Thanks for listening!

Questions?

