Project 16
Weekly Update: Jun 29, 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.

Dongmin Park[1†*], Minkyu Kim[1†*], Beongjun Choi[1†*], Junhyuck Kim[1†], Keon Lee[1†], Jonghyun Lee[1†], Inkyu Park[1†], Byeong-Uk Lee[1†], Jaeyoung Hwang[1†], Jaewoo Ahn[1,2†], Ameya S. Mahabaleshwarkar[3], Bilal Kartal[3], Pritam Biswas[3], Yoshi Suhara[3], Kangwook Lee[1,4,5], Jaewoong Cho[1]
[1]KRAFTON, [2]Seoul National University, [3]NVIDIA, [4]University of Wisconsin-Madison, [5]Ludo Robotics

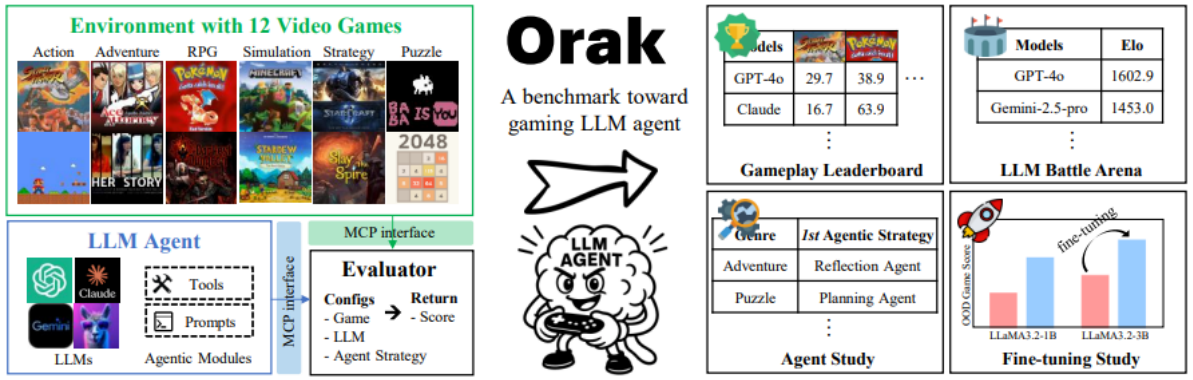
Orak (ICLR 2026)

Summary:

- Evaluates LLM/VLM performance on 12 existing video games

Our Contribution:

- We are creating our own game in a controlled environment
- We can understand reasons why the model failed with targeted questions



Seth Karten [*1] Andy Luu Nguyen [1] Chi Jin [1]
[1]Princeton University

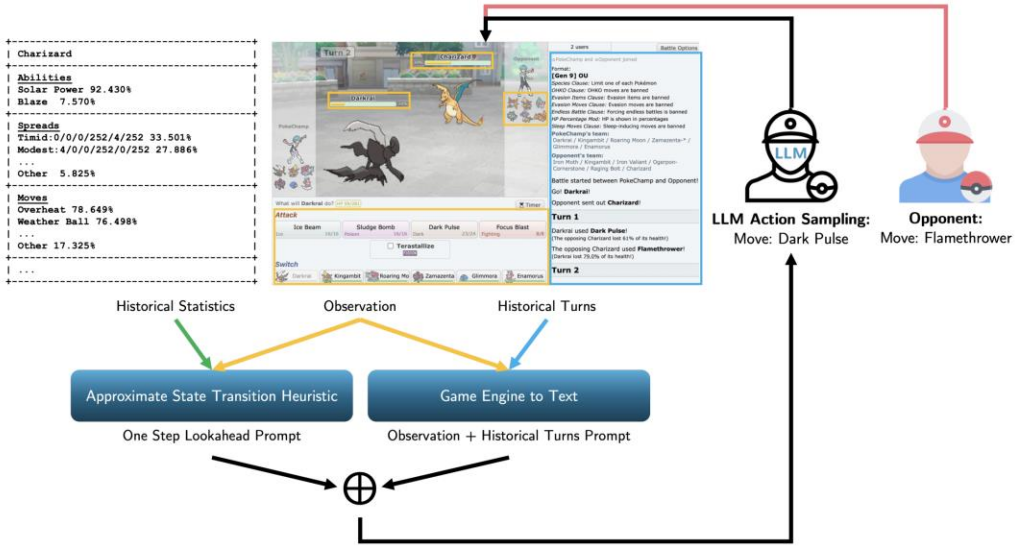
PokéChamp (ICML 2025)

Summary:

- Uses an LLM supported minimax search to choose the best move to win battles

Our Contribution:

- We are using VLMs to see the board state
- Our VLM is not being guided by a minimax search to play the game



Zhanyue Qin[1]✉, Haochuan Wang[1], Deyuan Liu[1], Ziyang Song[1], Cunhang Fan[2], Zhao Lv[2], Jinlin Wu[3], Zhen Lei[3], Zhiying Tu[1], Dianhui Chu[1], Xiaoyan Yu[3], Dianbo Sui[1]✉ *
[1] Harbin Institute of Technology
[2] Anhui University
[3] Institute of Automation, Chinese Academy of Sciences

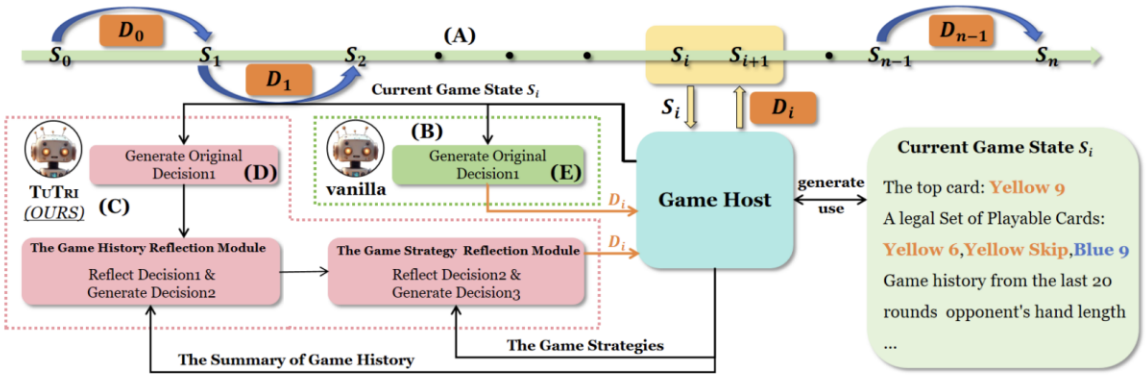
Uno Arena (EMNLP 2024)

Summary:

- Evaluates LLMs in a constructed UNO environment against heuristic players

Our Contribution:

- We are creating a much more complex game
- UNO arena is strictly text-based



Danrui Li[1], Sen Zhang[1], Samuel S. Sohn[1], Kaidong Hu[1], Muhammad Usman[2], and Mubbasir Kapadia[1,3]

[1]Rutgers University, NJ, USA
[2]Ontario Tech University, ON, Canada
[3]Roblox, San Meteo, CA, USA

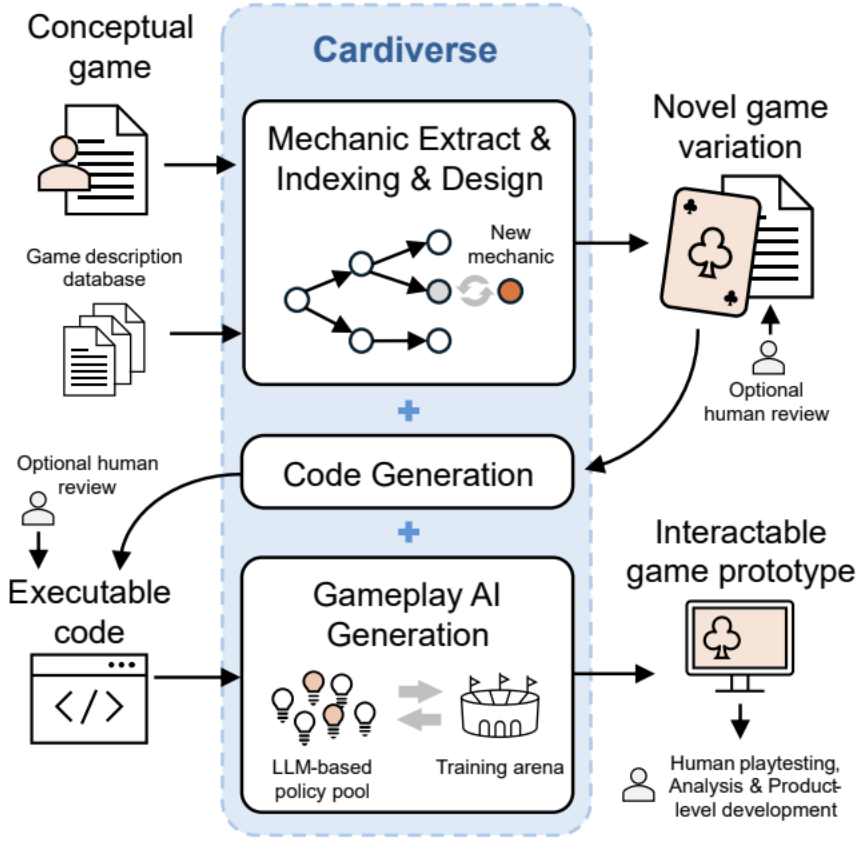
Cardiverse (EMNLP 2025)

Summary:

- Create card games and heuristic action-value functions using LLMs

Our Contribution:

- We are evaluating VLMs in how they understand the game, not as a development tool
- Cardiverse does not have LLM/VLMs actually playing the game, they just create bots



Jonathan Light[1]* Min Cai[2]* Sheng Shen[3] Ziniu Hu[4]

[1]Rensselaer Polytechnic Institute, [2]Shenzhen University
[3]University of California, Berkeley, [4]California Institute of Technology

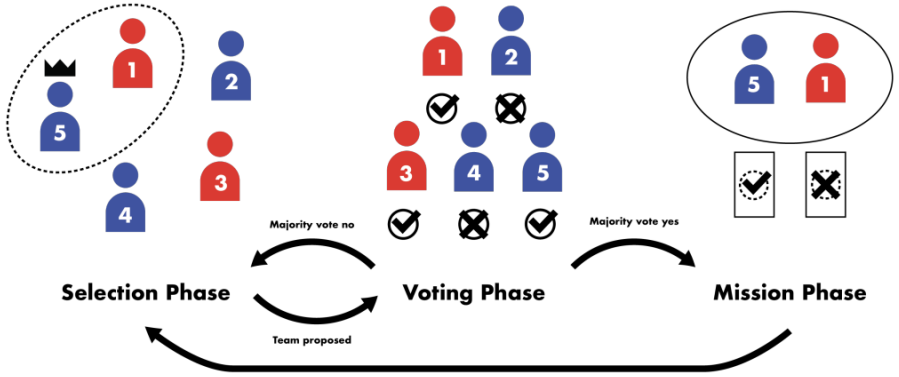
AvalonBench (NeurIPS 2023, Oral)

Summary:

- Evaluates LLMs in a social deduction game playing against rule-based bots

Our Contribution:

- Introduce stricter VLM evaluation
- We focus on visual game-state understanding, a different line of thinking



Minhua Lin[1] Enyan Dai[2] Hui Liu[3] Xianfeng Tang[3] Yuliang Yan[2] Zhenwei Dai[3]
Jingying Zeng[3] Zhiwei Zhang[1] Fali Wang[1] Hongcheng Gao[4] Chen Luo[2]
Xiang Zhang[1] Qi He[5] Suhang Wang[1]
[1] The Pennsylvania State University [2] HKUST (GZ) [3] Amazon
[4] Tsinghua University [5] Microsoft

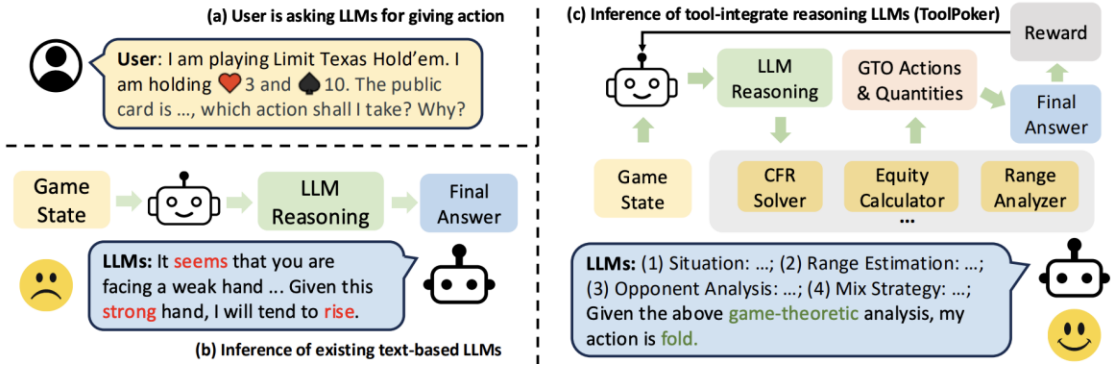
How Far Are LLMs from Professional Poker Players? (ICLR 2026)

Summary:

- Evaluates LLMs on how well they can play poker

Our Contribution:

- Poker has existing theory and solvers, our game is novel and more complicated
- We will use VLMs to analyze the board state



Richard Zhuang[1], Akshat Gupta[1*], Richard Yang[1], Aniket Rahane[1], Zhengyu Li[2], Gopala Anumanchipalli[1]
[1]University of California, Berkeley; [2]Georgia Institute of Technology

PokerBench (AAAI 2025)

Summary:

- Evaluates LLMs on distinct poker scenarios from a dataset and then fine-tune models
- Input: Text Description of River and Hand (**A♠ 6♥ J♦**)
- Output: Best action from current state

Our Contribution:

- We play entire games and evaluate many metrics such as win rate
- We judge how a VLM understands the visual state of the cards, not just a text input

Commonly Used Evaluation Metrics

- We will be using heuristic bots for evaluating LLM/VLM performance
  - Bots play against the model
  - Games are used to generate our dataset
- This is a commonly used strategy for developing LLM/VLM baseline
  - **Uno Arena**, **PokeChamp**, **AvalonBench**, and **Cardiverse** all use heuristic bots

Commonly Used Evaluation Metrics

- Evaluation Table from PokeChamp paper (ICML 25)

Table 2: Performance in Gen 9 OU battles with terastallize mechanic and custom teams.

Method	LLM	Win Rate vs. Abyssal (%)	Elo	Avg. # Turns
PokéChamp	GPT-4o	**84**	**1268**	**15.7**
PokéChamp	Llama 3.1:8b	56	1204	16.9
PokéLLMon	GPT-4o	40	1020	22.6
Abyssal	N/A	N/A	1117	17.9
One Step Lookahead	N/A	44	1107	17.9
Max Power	N/A	16	885	19.5
Random	N/A	0	399	21.2

Discussion on Pre-Built Bots

- Humans are better than the developed bots
  - Sarinda vs Random: 100% win-rate
  - Sarinda vs Rule-Based: 100% win-rate
  - Sarinda vs Greedy: 100% win-rate
- BUT, bots are currently better than the LLMs/VLMs
- **Result:** The bots are a useful intermediate evaluation to see if the models can play the game
  - If a model beats a bot more often, it would also improve against a human player

Model vs Bot (48 Games per matchup)

Win Percentage				
	Random	**Rule-Based**	**Greedy**	**Overall***
QwenLLM-32B	83.3%	10.4%	16.7%	36.80%
Llama-8B	29.2%	4.2%	8.3%	13.90%
LLaVA-7B	**91.7%**	**15.2%**	**27.1%**	**44.67%**
InternVL-38B	85.4%	14.6%	0.0%	33.33%
QwenVLM-32B	91.5%	10.9%	15.9%	39.43%
Sarinda*	***100%***	***100%***	***100%***	***100%***

Weeks 6-7 Deliverables

Task	Week 6-7
Share and Play	🟡 (40/100 games)
Generate larger dataset for game understanding questions	🟡
Complete evaluation on larger dataset	🟡
Evaluate model vs MCTS	🟡 (Running)
Play language models against each other	🟡
Implement memory module using collected data	🟡
Literature review on memory	✅

Thanks for listening!

Questions?

