Reasoning Under Pressure: LLMs in Competitive Pokémon Battles

Tadisetty Sai Yashwanth
Turilabs
Bangalore, India
taddishetty34@gmail.com

Dhatri C
Turilabs
Bangalore, India
dhatri.c22@gmail.com

Abstract

We introduce LLM Pokémon League, a system that uses competitive Pokémon battles to study how large language models (LLMs) reason and make strategic decisions. In this setup, models from Openai, Anthropic, and Google face each other in tournaments where they must build teams, choose moves, and adapt to uncertain situations. Each action is explained in natural language, allowing us to closely examine how models think, adjust, and plan during the game. Unlike other evaluation methods that require large resources, Pokémon League offers a lightweight and accessible way to see real-time strategy and reasoning. Our experiments show clear differences in how models approach battles, from careful team balance to risk-heavy play styles. By framing reasoning as a competitive game with transparent choices, LLM Pokémon League provides a practical way to compare and understand the strategic abilities of today's leading models.

1 Introduction

Large reasoning models (LRMs) have recently made great progress in tasks like planning, multi-step problem solving, and symbolic reasoning. But most ways of testing these abilities, such as math problems, code generation, or logic puzzles do not fully capture how models make decisions in dynamic and uncertain situations. These settings show reasoning ability in structured domains, but they miss the kind of adaptability and tactical trade-offs that are central to real decision making.

To address this, we introduce LLM Pokémon League, a system that places models in competitive Pokémon battles. Unlike static problem-solving tasks, battles require models to act under pressure, adapt to hidden information, and balance short-term tactics with long-term strategy. In a battle setting, every move matters, and we can see how models respond in "pinch" situations where one wrong step may cost the game. This makes it a natural testbed for evaluating intelligence through action, adaptation, and strategy.

In our framework, models from OpenAI, Anthropic, and Google act as Pokémon "trainers" in tournaments. Before each match, they must select a team of Pokémon, balancing type coverage, roles, and potential counters. During battles, they then decide whether to attack, switch, or plan ahead, while also explaining their reasoning in natural language. This setup allows us to study both outcomes such as win rates and move efficiency, and reasoning processes such as consistency, adaptability, and risk-taking.

By framing reasoning as competitive play, LLM Pokémon League highlights how different models plan, adapt, and strategize under adversarial conditions. Our results show convergence in some areas, such as favoring balanced teams, but also reveal divergent styles, from cautious and incremental play to aggressive, high-risk strategies.

Submitted to NeurIPS 2025 Workshop on Efficient Reasoning. Do not distribute.

All code, battle logs, and evaluation scripts are publicly available at https://github.com/Turi-Labs/Poke-Bench

2 Related Works

Prior evaluations of LLM reasoning have largely focused on structured domains such as mathematics, code generation, and symbolic logic, where performance is measured by accuracy under chain-of-thought or extended context prompts. While effective, these approaches are often compute-intensive and reveal little about reasoning in adversarial or interactive environments.

Work in multi-agent and game-based contexts has shown both promise and limitations. Behavioral game theory studies [1] and behavioral economics games [2] reveal that models display some strategic tendencies but struggle with higher-order reasoning. Recursive reasoning frameworks [3] and multi-agent prompting [4] highlight how distributing reasoning can enhance performance on complex tasks, motivating further exploration of competitive multi-agent environments.

Pokémon has recently emerged as a useful testbed for adversarial reasoning. Research includes supervised fine-tuned agents [7], domain-specific benchmarks such as PokeLLMon [8], and competitions emphasizing win rates [9]. Other directions integrate symbolic reasoning with reinforcement learning for deliberative alignment [5] or emphasize tracing decision rationales in strategic games [6]. While these efforts demonstrate the value of such settings, they often rely on handcrafted heuristics or lack transparency into models' reasoning.

In contrast, our framework treats Pokémon as a lightweight, general-purpose reasoning benchmark. By evaluating models in zero-shot conditions and capturing natural-language rationales alongside actions, we focus not only on outcomes but also on interpretability, addressing gaps in resource-efficient, transparent benchmarks for strategic reasoning.

3 Methodology and Setup

The LLM Pokémon League is designed as a lightweight, easy-to-understand environment for evaluating the strategic reasoning of large language models (LLMs) in adversarial multi-agent settings. It consists of three key components: the tournament setup, the interaction pipeline, and the reasoning capture process.

3.1 Tournament Setup

We set up multiple single-elimination bracket of eight competing agents, each backed by a state-of-the-art LLM. Models evaluated include OpenAI GPT-4.1, o4-mini, o3; Anthropic Claude Sonnet 3.5, 3.7, 4; and Google Gemini 2.5 Pro and Flash. All models are used in a zero-shot setting without domain-specific fine-tuning, preserving their general-purpose reasoning ability. Tournament pairings are initialized with a random seed to avoid bias in matchups and ensure robustness across repeated runs.

Prior to each match, agents perform team selection from a curated pool of 60 Pokémon spanning Generations I–III. The pool is balanced for type coverage and strategic diversity. Models are instructed to select six Pokémon while explaining their rationale in natural language. This phase emphasizes multi-objective reasoning under constraints, such as balancing weaknesses, coverage, and synergy.

3.2 Interaction Pipeline

Battles are conducted in a turn-based environment with structured prompts describing the current game state: active Pokémon, health/status, known opponent attributes, and available moves or switches. Models return both an action (attack or switch) and a natural language explanation of their choice.

Example:

```
Prompt:  Your active Pokémon:  Jolteon (72% HP). Opponent's
Pokémon:  Gyarados (60% HP). What will you do?
Response:  "I will use Thunderbolt because Gyarados is
Water/Flying, weak to Electric.  Jolteon outspeeds, so this
is the safest option."
```

Actions are executed via the battle engine written in python, which resolves damage, effects, and win/loss conditions. The match proceeds turn by turn until one player's entire team of six Pokémon has fainted, producing a complete decision log with rationales for every move.

3.3 Reasoning Capture

All model outputs are stored in a structured JSON schema that records both the chosen actions and the accompanying free-text justifications. This pairing allows us to examine not only what decision was made, but also why the model believed it was the appropriate choice. Such a design enables fine-grained analysis across several dimensions of reasoning. For instance, we can assess reasoning–action alignment by checking whether the explanation matches the executed move, investigate evidence of opponent modeling when a model anticipates potential switches or counters, and evaluate risk management in cases where a model chooses to pivot defensively or sacrifice a weakened Pokémon for strategic advantage.

By systematically linking actions to rationales, the framework ensures that every decision can be interpreted in context, building on prior work in multi-stage game reasoning [6] while providing a more transparent lens into LLM decision-making.

4 Results and Analysis

The results of the LLM Pokémon League reveal not just which models won or lost, but how they reasoned their way through strategic challenges. Rather than reducing performance to win rates alone, we analyzed recurring heuristics, tactical adaptations, and divergent play styles across multiple tournaments. This uncovers consistent reasoning patterns that shaped victories, from cautious risk management to aggressive exploitation of statistical advantages. These findings illustrate that even within the same constrained environment, LLMs can develop distinct, interpretable approaches to winning. In the following subsections, we detail the specific patterns observed, beginning with team formation strategies.

4.1 Team Selection Patterns

Across tournaments, clear convergent heuristics emerged in how models constructed their teams. Nearly all agents prioritized **broad type coverage**, carefully avoiding redundant weaknesses while ensuring offensive reach across multiple matchups. Most also gravitated toward balanced compositions that mirrored competitive human play, combining **offensive sweepers** such as Salamence and Rayquaza with **defensive anchors** like Swampert and Blissey, supported by versatile pivots such as Skarmory. A recurring trend was the frequent inclusion of Steel-types like Metagross and Skarmory, which functioned as universal counters, an indication that models were not only optimizing for immediate strength but also **anticipating** likely opponent choices.

Yet within this broad convergence, notable divergences appeared. Some models pursued **unconventional strategies** that reflected distinct reasoning styles. For example, `o4-mini` constructed a roster stacked with legendaries (Kyogre, Groudon, Rayquaza, Ho-Oh, Lugia), leveraging their overwhelming base stats and synergistic weather effects to overwhelm opponents in high-variance battles. In contrast, `o3` consistently favored a conventional balance core, built around Swampert, Zapdos, Metagross, and Blissey, prioritizing **stability** and **long-term resilience** over raw power. These differences highlight how LLMs, even under identical constraints, can arrive at contrasting strategic philosophies, ranging from risk-heavy dominance plays to measured, methodical team design.

4.2 In-Battle Reasoning

Once battles began, models consistently displayed tactical reasoning that was both interpretable and strategically sound. Across matches, they demonstrated reliable mastery of the type chart, prioritizing **super-effective** attacks and avoiding low-value moves. They also showed a tendency to **preserve high-value Pokémon** at low HP, echoing late-game resource management strategies often employed by human players.

When faced with a choice between risk and consistency, models frequently favored **accurate and reliable moves**. For example, opting for Thunderbolt rather than the riskier but stronger Thunder. Even under disadvantageous matchups, they often sought to mitigate losses by pivoting into neutral absorbers rather than carelessly sacrificing resources.

Table 1: Aggregate Tournament Results across 10 runs

Model	Tournaments Won	Overall Strategy
o4-mini	6	Dominated with legendary-heavy teams
o3	3	Relied on balanced cores
claude-sonnet-4	1	Preferred defensive, counter-based selections
gpt-4.1	0	Favored accurate, low-variance moves
claude-3.5-sonnet	0	Tended toward simple, low-variance teams
claude-3.7-sonnet	0	Similar to 3.5: conservative picks
gemini-2.5-pro	0	Often built one-dimensional or unbalanced teams
gemini-2.5-flash	0	Erratic, low-consistency play

Taken together, these behaviors suggest that LLMs are not merely recalling isolated facts but are actively leveraging pretrained general knowledge to navigate novel, dynamic states. Their decision-making reflects an ability to balance immediate tactical gain with long-term resource conservation, a hallmark of strategic reasoning under pressure.

4.3 Tournament Outcomes

Across ten tournaments, clear performance hierarchies emerged. o4-mini established itself as the dominant agent, winning six of the ten brackets. Its success stemmed from consistently executing a high-risk, high-reward strategy centered on legendary-heavy rosters, which overwhelmed more conservative opponents. The next strongest performer, o3, secured three tournament victories by relying on balance cores that emphasized stability and coverage, reflecting a more methodical reasoning style. Sonnet-4 managed a single tournament win, typically advancing deep but struggling to overcome the sheer force of o4-mini.

Other models showed mixed or consistently weak results. GPT-4.1 was notably reliable, reaching the semifinals in nearly every run and only suffering an unexpected upset once against Gemini Flash 2.5. By contrast, Claude 3.5 and Claude 3.7 rarely advanced, securing only two wins across all tournaments combined. The weakest overall performer was Gemini Flash, which lost most of its matches decisively, with the exception of a single upset victory over GPT-4.1.

These outcomes, summarized in Table 1, highlight the divergence between dominant strategies and those that failed to generalize. Importantly, the repeated success of o4-mini demonstrates that even under identical conditions, some models are more willing to embrace variance-heavy strategies, while others default to conservative balance, and this divergence in “personality” directly shaped competitive outcomes.

5 Discussion and Conclusion

Our results highlight three key insights: LLMs demonstrate layered reasoning and adaptive planning even in compact domains; most converge on balanced, human-like strategies, yet outlier approaches such as o4-mini’s legendary-heavy teams proved decisive. Natural-language rationales provide transparency into both successes and failures, enabling analyses beyond opaque benchmarks.

The LLM Pokémon League thus offers a lightweight yet strategically rich testbed for studying reasoning under pressure. It surfaces trade-offs between efficiency, adaptability, and creativity without the heavy computational demands of traditional benchmarks. Looking ahead, scaling to larger tournaments, testing smaller model variants, and extending to other adversarial domains could further validate this approach.

Compact competitive environments can yield rich insights into model reasoning, offering an interpretable and resource-efficient complement to existing evaluation methods.

References

[1] Jia, Y., et al. (2025). Disentangling Reasoning Ability and Contextual Effects in Large Language Models via Behavioral Game Theory. *Preprint*.

[2] Lee, S., & Kader, A. (2024). Evaluating Strategic Reasoning of Large Language Models in Behavioral Economics Games. *AAAI Conference on Artificial Intelligence*.

[3] Zhang, H., et al. (2025). K-Level Reasoning: Recursive Theory of Mind in Large Language Models. *NeurIPS*.

[4] Anthropic. (2025). How We Built Our Multi-Agent Research System. *Anthropic Engineering Blog*.

[5] Guan, X., et al. (2024). Deliberative Alignment: Enhancing Safety and Robustness in Large Language Models through Symbolic Reasoning and Reinforcement Learning. *Technical Report*.

[6] Yuan, X., et al. (2025). Tracing LLM Reasoning Processes with Strategic Games: A Framework for Planning, Revision, and Resource-Constrained Decision Making. *arXiv preprint arXiv:2506.12012*.

[7] Lv, Z., & Qihang, C. (2024). Pokémon Battle Agent based on LLMs. *THU Fall AML Submission*

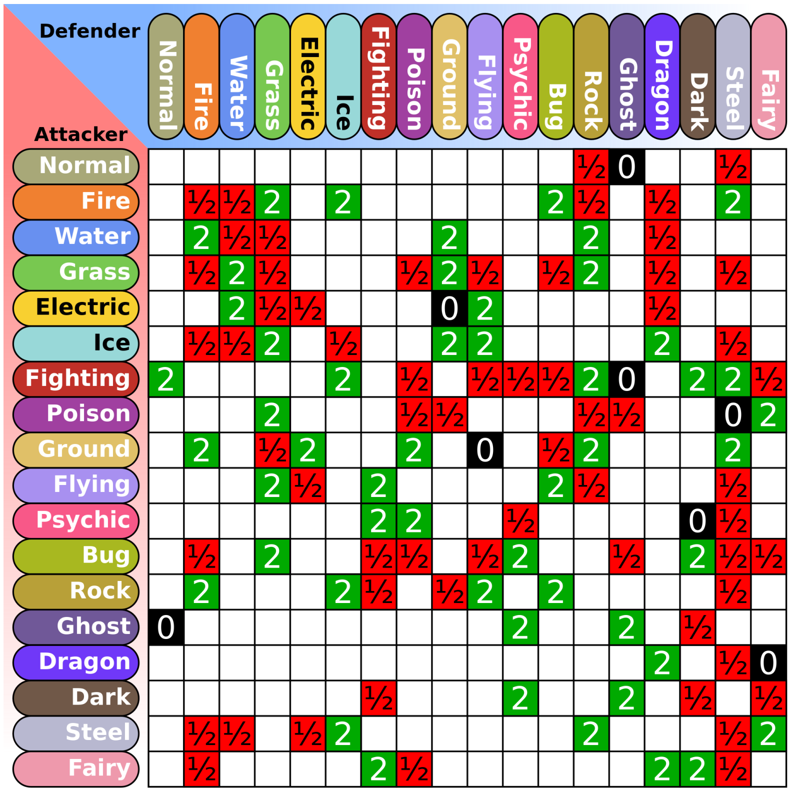
[8] Hu, S., Huang, T., Liu, G., Kompella, R. R., & Liu, L. (2025). PokéLLMon: A Grounding and Reasoning Benchmark for Large Language Models in Adversarial Pokémon Battles. *ICLR Workshop*.

[9] VGC AI Competition. (2025). IEEE Conference on Games (CoG 2025) AI Competition. https://cog2025.inesc-id.pt/vgc-ai-competition/

A Appendix

A.1 Why Pokémon Battles are a Good Test of Reasoning

As shown in Figure 1, the Pokémon type effectiveness chart illustrates damage multipliers among 18 different types. Mastery of this system is critical for strategic play and underpins many of the models' tactical choices.



Defender Attacker	Normal	Fire	Water	Grass	Electric	Ice	Fighting	Poison	Ground	Flying	Psychic	Bug	Rock	Ghost	Dragon	Dark	Steel	Fairy
Normal													½	0			½	
Fire		½	½	2		2						2	½		½		2	
Water		2	½	½					2				2		½			
Grass		½	2	½				½	2	½		½	2		½		½	
Electric			2	½	½				0	2					½			
Ice		½	½	2		½			2	2					2		½	
Fighting	2					2		½		½	½	½	2	0		2	2	½
Poison				2				½	½				½	½			0	2
Ground		2		½	2			2		0		½	2				2	
Flying				2	½		2					2	½				½	
Psychic							2	2			½					0	½	
Bug		½		2			½	½		½	2			½		2	½	½
Rock		2				2	½		½	2		2					½	
Ghost	0										2			2		½		
Dragon															2		½	0
Dark							½				2			2		½		½
Steel		½	½		½	2							2				½	2
Fairy		½					2	½							2	2	½	

Figure 1: Pokémon type effectiveness chart, illustrating damage multipliers among 18 different types.

Pokémon battles provide a uniquely rich environment for testing strategic reasoning in LLMs. Unlike purely symbolic benchmarks, battles require agents to balance multiple layers of decision-making:

- **Type Matchups:** With 18 Pokémon types and over 300 type interactions, every move involves combinatorial reasoning about effectiveness (e.g., Water beats Fire, but loses to Grass). Models must recall these relationships and apply them in context.
- **Resource Management:** Limited health points, finite move power points (PP), and status effects (paralysis, sleep, poison) force trade-offs between immediate gains and long-term preservation.
- **Partial Information:** Agents cannot see the opponent's full roster or moveset at the start, requiring prediction and opponent modeling.
- **Multi-step Planning:** Decisions extend beyond the immediate turn—sacrificing a weakened Pokémon can set up a sweep for another teammate, similar to sacrificing in chess.

These factors combine into a compact yet strategically deep environment. Unlike math or code benchmarks, Pokémon provides an interpretable, adversarial setting where reasoning can be studied both in terms of outcomes (win/loss) and process (natural-language rationales).

A.2 Extended Insights from Tournament Play

As shown in Figure 2, some Pokémon were consistently favored by multiple models, while others appeared only in divergent strategies.

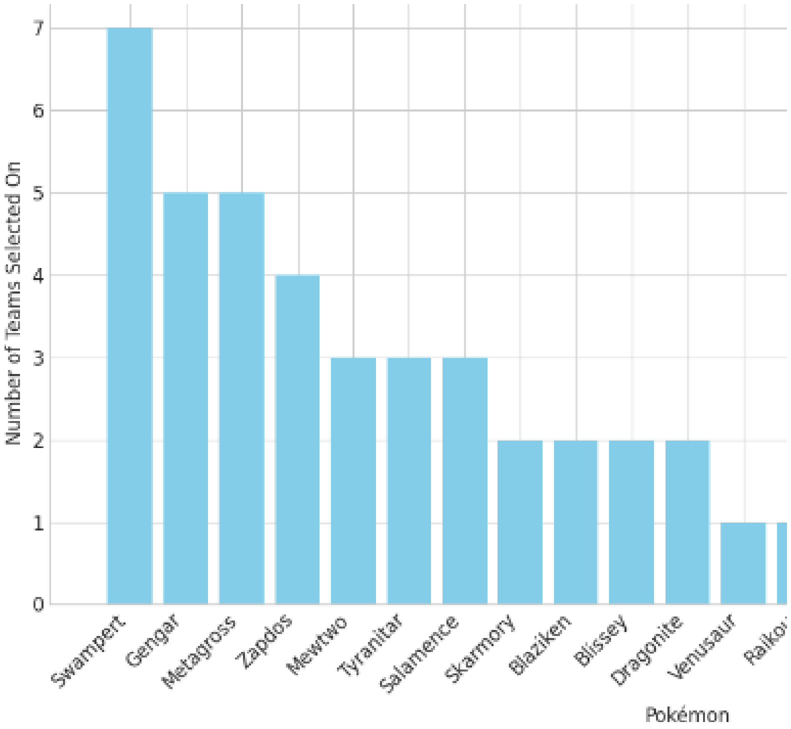


Figure 2: Frequency of Pokémon selection across all 8 LLM agent teams. High pick rates for Swampert (6/8) and Metagross (5/8) reflect convergent recognition of strong typing and utility. The champion's team, however, leaned heavily on legendary picks like Kyogre and Groudon.

The full single-elimination tournament (8 agents, 7 matches) highlighted several emergent reasoning patterns:

- **Convergence and Divergence:** Six of eight models independently selected Swampert, and five picked Metagross, reflecting convergent recognition of their strong defensive utility. In contrast, the champion (`o4-mini`) uniquely stacked legendary Pokémon with weather synergy, showing divergent but successful strategy.

- **Risk Profiles:** Some models (e.g., o3, gpt-4.1) consistently preferred accurate and conservative moves (Thunderbolt over Thunder), while others (o4-mini) embraced higher-reward lines, leading to more decisive outcomes.
- **Opponent Modeling:** Across matches, agents made predictive pivots (e.g., switching Salamence to absorb Earthquake with Intimidate) that suggest non-myopic reasoning rather than one-step lookahead.
- **Faithfulness Gaps:** Rationales sometimes diverged from the actual game state, such as o4-mini misidentifying Gengar's Ground immunity. These contradictions highlight the importance of evaluating not only decisions but also rationale alignment.

Taken together, these observations suggest that Pokémon battles naturally elicit reasoning behaviours spanning planning, adaptation, and error. They provide both a competitive outcome signal (win/loss) and interpretable rationale traces, making them a useful complement to traditional reasoning benchmarks.

A.3 Reasoning Length per Match

To connect tournament play with efficiency, we tracked the length of reasoning each model produced per match. Table 2 summarizes average words and characters per match.

Table 2: Reasoning Length per Match

Model	**Words**	**Characters**
o4-mini	664	4,283
o3	1,770	10,991
gpt-4.1	1,816	11,223
claude-sonnet-4	1,189	7,636
claude-3.7-sonnet	2,623	16,743
claude-3.5-sonnet	1,963	12,878
gemini-2.5-pro	1,901	11,800
gemini-2.5-flash	1,459	9,829

Interestingly, more verbose models did not consistently achieve better outcomes. For instance, claude-3.7-sonnet generated the longest rationales but was often eliminated early, while o4-mini combined concise reasoning with decisive victories. This suggests that reasoning quality (faithfulness and conciseness) may matter more than sheer verbosity.