From Rules to Nash Equilibria: Formally Verified Game-Theoretic Analysis of a Competitive Trading Card Game

Author names withheld for review

Abstract—We present a formally verified analysis of competitive Pokémon Trading Card Game metagame dynamics using Lean 4 and real tournament data. Our formal development spans approximately 30,000 lines across 75 Lean files and proves over 2,500 theorems without sorry, admit, or custom axioms. On top of this foundation, we analyze Trainer Hill tournament data from January 29 to February 19, 2026 (50+ player events), covering 14 archetypes and a complete 14×14 matchup matrix. The central result is a verified popularity paradox: Dragapult Dusknoir is the most played deck (15.5%) but has only 46.7% expected win rate against the field, while Grimmsnarl Froslass (5.1% share) has the highest expected field win rate at 52.7%. We further show an extreme mixed-strategy Nash profile concentrated on Mega Absol Box (∼93%) and Dragapult (∼7%), prove replicator dynamics pressure against Dragapult and Ceruledge, and quantify best-of-three amplification effects such as 67.3% → 74.9% for Raging Bolt Ogerpon versus Mega Absol Box. These results demonstrate that machine-checked formal methods can produce empirically grounded and strategically actionable conclusions in modern competitive game ecosystems.

Index Terms—Formal verification, game theory, trading card games, Nash equilibrium, theorem proving, metagame analysis, replicator dynamics, Lean 4

I. Introduction

Tournament outcomes in competitive trading card games (TCGs) are often shaped before round one begins. Players must make two coupled decisions: how to play each game state and which deck to register. The first decision is local and tactical; the second is global and game-theoretic. The pre-tournament deck-selection problem is naturally modeled as a strategic game where payoffs derive from matchup win rates and the population distribution of opponents.

The Pokémon TCG is especially suitable for this analysis. It has a large organized-play ecosystem, clearly defined public rules, and a metagame that evolves quickly enough to produce measurable strategic cycles. At the same time, the domain is difficult for informal reasoning: hidden information, stochastic effects, and nonlinear tournament incentives make intuition unreliable even for experienced players.

Formal methods offer an attractive remedy. By encoding game semantics in Lean 4 [1] and proving strategic statements directly over exact data representations, we separate factual claims from narrative claims. A statement in the paper is either derivable from formally checked definitions and theorems or it is excluded. We build a proof-carrying metagame analytics pipeline where the verified objects are (i) data representation and ingestion, (ii) expected-value computations over the field, (iii) Nash equilibrium and replicator dynamics statements, and (iv) tournament-objective transforms (Bo3, Swiss). The in-game rules formalization serves as supporting infrastructure guaranteeing legality and enabling future counterfactual analysis.

Our empirical foundation is Trainer Hill metagame data, aggregated from Limitless tournament records [2], [3], for 50+ player tournaments from January 29 to February 19, 2026. Within this fixed window, we model the top 14 archetypes and their full pairwise matchup matrix. The resulting matrix contains enough granularity to support rigorous expected-value computations, equilibrium analysis, and evolutionary dynamics without introducing synthetic assumptions about win rates.

This paper makes four concrete contributions aligned with this thesis. First, we formalize the rules and legality substrate that supports trustworthy ingestion and future counterfactual analysis. Second, we encode real metagame data as exact values and verify expected-value computations, including the popularity paradox. Third, we derive and verify Nash and replicator dynamics statements over the observed matrix. Fourth, we verify tournament-objective transforms from single-game payoffs to best-of-three and Swiss-relevant decision criteria.

A key practical implication follows immediately. Popularity is not a proxy for optimality. The most visible deck in the room may be strategically dominated once weighted against the actual field. Formal verification provides a robust way to expose this mismatch and quantify its consequences for player choice.

The remainder of the paper is organized as follows. Section II situates the work. Section III presents the Lean model of rules and legality. Section IV develops probability and resource theory. Section V details data and measurement. Section VI presents the popularity paradox. Section VII analyzes equilibrium and dynamics. Section VIII discusses tournament strategy. Section IX documents formalization methodology. Section X covers validity threats, and Section XI concludes.

A. Motivating Tournament Scenario

Consider a player preparing for a 10-round open tournament. Community sentiment says Dragapult is “the deck to beat” because it is popular, has many practiced pilots, and appears frequently in streamed matches. An informal decision process often stops here: players mirror the visible metagame and assume that high adoption signals high objective value.

Our data and proofs show why this shortcut fails. Popularity is an endogenous variable shaped by familiarity, deck cost, and social diffusion. Expected win rate, in contrast, is a payoff object determined by pairwise matchups weighted by opponent frequencies. These are related but not equivalent quantities.

Suppose the same player has two candidate decks: Dragapult and Grimmsnarl. The first has better perceived comfort and social validation; the second has lower adoption but stronger weighted performance in this snapshot. Without a formal model, one may over-weight convenience and under-weight aggregate matchup structure. With a formal model, the decision can be recast as a transparent optimization problem with explicit assumptions.

This distinction scales from individuals to populations. If many players independently adopt “safe” visible decks despite negative expected fitness, the field can stabilize in a collectively suboptimal state. That state is exactly what our popularity paradox theorem captures. In this sense, the theorem is not merely descriptive; it is diagnostic of strategic inefficiency in ecosystem-level behavior.

The practical value for competitors is immediate. A formally validated metagame model provides a defensible framework for deciding when to exploit, when to hedge, and when to avoid overreacting to noisy week-to-week narratives. For researchers, it provides a rare domain where rich empirical data and theorem-prover rigor can be combined in one reproducible pipeline.

II. Related Work

A. Formal Methods and Strategic Games

Formal reasoning has transformed analysis in several strategic domains. In classical games, foundational communication and search analyses and later complexity results for generalized chess established the computational stakes of strategic reasoning [4]–[6]. In imperfect-information settings such as poker, game-theoretic systems like Cepheus, Libratus, and Pluribus [7]–[9] show that equilibrium reasoning can scale when abstractions are carefully managed. At larger multi-agent scale, AlphaZero and AlphaStar demonstrate superhuman play in complex domains [10], [11].

TCGs are harder in a different way. Their action spaces and card interactions are highly compositional, and rules often involve exception-heavy textual semantics. This combination increases the risk of silent modeling errors when analysis is implemented in ad hoc scripts. A proof assistant mitigates this risk by forcing explicit treatment of definitions, invariants, and edge cases.

B. AI and Metagame Analysis in Card Games

Prior card-game AI work has emphasized in-game decision quality, including Monte Carlo methods for games such as Magic and Hearthstone [12]–[15]. Related work also includes deckbuilding optimization and competition environments for TCG AI [16]–[18]. That line of work is valuable but orthogonal to the question we study: how should a player choose a deck before round one, given a population distribution and matchup matrix?

Metagame-level analysis appears frequently in practitioner content but rarely as reproducible formal scholarship. The missing ingredients are a fixed data definition, a mathematically explicit payoff model, and mechanically checked claims. Our approach fills this gap by treating metagame analysis as theorem proving over empirical constants.

C. Theorem Proving for Rule Systems

Lean 4 [1] has become a practical environment for large-scale formalization because it combines expressive dependent types with efficient decision procedures. Large collaborative libraries and landmark formal proofs further show the maturity of this ecosystem [19]–[22]. For game-rule modeling, dependent types are particularly useful: constraints such as bounded bench size, deck legality predicates, and well-formed transitions can be encoded directly as propositions attached to data.

Closest in spirit is prior formalization work on card-game effects in Isabelle/HOL [23], which established feasibility. Our work differs by coupling rule formalization to a complete, real matchup matrix and then pushing through equilibrium and dynamics claims tied to observed tournament distributions.

D. Evolutionary and Behavioral Perspectives

Replicator dynamics [24]–[26] provide a natural lens for metagame adaptation: strategies with above-average fitness gain share, while below-average strategies lose share. In practice, observed metagames often drift slowly because human behavior is not perfectly rational. This motivates interpreting deviations from equilibrium in behavioral-economic terms rather than treating them as model failure.

The contribution here is methodological: we use evolutionary tools not as informal metaphors but as theorem-backed statements over fixed data. When we claim that Dragapult has negative relative fitness or that Ceruledge declines under replicator updates, those claims are machine-checked consequences of the encoded matrix and shares.

E. Community Analytics Versus Proof-Carrying Analytics

Competitive communities already produce large volumes of metagame commentary. Those outputs are valuable for speed, but they typically combine raw percentages, subjective confidence, and hand-waved matchup transitivity. This workflow is well suited for rapid iteration and poor at preventing silent arithmetic or modeling mistakes.

Our approach is intentionally stricter. Every quantity referenced in argument is represented as an explicit program term. Every strategic claim is either the direct output of a computation over those terms or a theorem whose proof is validated by the Lean kernel. The cost is additional formalization overhead; the benefit is auditability and long-term reproducibility.

This positioning is complementary rather than adversarial. Community analytics can generate hypotheses quickly. Formal analytics can then validate, reject, or refine those hypotheses with machine-checkable guarantees. In practice, the two pipelines can coexist: empirical dashboards provide breadth, while theorem-prover workflows provide depth and reliability for high-impact claims.

The popularity paradox is a good example. Informally, one might suspect overplay from anecdotal matchup frustration. Formally, we can prove the exact inequality against normalized field weights and make all assumptions explicit. That transformation from intuition to theorem is the central methodological gap this paper addresses. Section III now defines the formal rule substrate used by the subsequent empirical and strategic analyses.

III. Game Formalization

We formalize the strategic layer of the Pokémon TCG by combining operational state semantics with legality and resource invariants, grounded in official rule documents [27], [28]. In this paper, the rules layer is supporting infrastructure rather than the primary empirical claim object for the 2026 snapshot. It future-proofs the framework for counterfactual analysis (e.g., "what if a card is banned?"), guarantees data-ingestion correctness through legality alignment, and enables mechanical derivation of matchup implications from legal state transitions. The 2026 snapshot analysis itself relies primarily on matrix-level verification over observed matchup data. The goal is not to mechanize every card text in existence; instead, we encode the rule substrate needed to reason about deck legality, turn progression, card flow, and payoff-relevant mechanics.

A. Game State Representation

The state model explicitly tracks per-player zones, turn ownership, and phase. This representation is sufficient to express legality constraints and to prove conservation and progress properties.

```
structure GameState where
  activeP1   : Pokemon
  benchP1    : List Pokemon
  handP1     : List Card
  deckP1     : List Card
  prizesP1   : List Card
  discardP1  : List Card
  activeP2   : Pokemon
  benchP2    : List Pokemon
  handP2     : List Card
  deckP2     : List Card
  prizesP2   : List Card
  discardP2  : List Card
  turnPlayer : Player
  turnPhase  : Phase
```

A state-level encoding gives us a uniform target for rules, tactics, and metatheory. In particular, all strategic claims can be connected to primitive transition semantics, preventing the common mismatch where high-level analysis assumes mechanics that are absent from the underlying rules implementation.

B. Turn Phases and Transition Discipline

Turn order is represented as a finite phase machine. This makes "what can happen next" decidable, enabling automation for many local proofs.

```
inductive Phase where
  | DrawPhase
  | MainPhase
  | AttackPhase
  | BetweenTurns
  deriving DecidableEq, Repr
```

By constraining transitions through this type, we can prove phase safety lemmas (e.g., no attacks during draw phase) and derive stronger progress claims. These lemmas are not just implementation detail: they ensure that all strategy-level simulations are grounded in legal game trajectories.

C. Type Effectiveness

Weakness and resistance are encoded as total functions over enumerated types. The basic Fire>Grass relation appears as a direct theorem.

```
theorem fire_beats_grass :
    weakness .Grass = .Fire := by
  rfl
```

Even elementary facts matter because they serve as trusted building blocks for larger proofs, especially when computing expected damage and trade sequences. Encoding them in the core logic avoids accidental divergence between prose assumptions and executable semantics.

D. Card Conservation and Trainer Effects

We model high-impact trainer cards with explicit zone transitions. For Professor's Research, we prove that discarding the hand and drawing seven preserves global card count across all zones.

```
theorem professors_research_conserves_cards
    (s s' : GameState) :
    stepProfessorResearch s = some s' ->
    totalCardCount s' = totalCardCount s := by
  intro h
  cases h
  native_decide
```

Conservation theorems are critical for trust. Without them, subtle bookkeeping bugs can distort probability estimates and strategic value calculations. With them, resource-theoretic statements in later sections inherit mechanical guarantees.

E. Deck Legality as a Biconditional

Deck legality is implemented as a computable checker and linked to an inductive specification through a soundness-and-completeness theorem.

```
theorem deckLegal_sound_complete (d : Deck) :
    checkDeckLegal d = true <-> DeckLegal d := by
  constructor <;> intro h <;> simp_all
```

This biconditional is essential for reproducibility. It guarantees that data ingestion and deck filtering in empirical analysis are extensionally equivalent to the formal legality policy, rather than "close enough" approximations.

F. Invariant Catalog and Proof Obligations

Beyond the highlighted theorems, the formal model maintains a broad invariant catalog used by downstream analysis and simulation tooling. Representative invariants include: (i) nonnegative zone sizes, (ii) deck-size preservation except under explicit draw/mill transitions, (iii) bench-size upper bounds, (iv) uniqueness constraints for once-per-turn actions, (v) deterministic transition behavior under fixed randomness traces, (vi) legality of retreat and switching operations, (vii) prize-card accounting invariants, (viii) terminal-state exclusivity conditions, and (ix) well-typed status-effect transitions.

These invariants matter for metagame work because matchup payoffs are aggregate outputs of many local game interactions. If local rule mechanics leak cards, skip phase guards, or violate zone consistency, macro-level expected values become unreliable. The invariant layer prevents this by making such inconsistencies theorem-level failures.

A second benefit is maintainability. As card-effect libraries expand, invariant checks serve as regression barriers. New effect encodings must satisfy shared structural properties before they can influence strategic analysis. This minimizes accidental model drift and keeps historical comparisons meaningful across commits.

Finally, invariants support modularity. Game semantics, probability modules, and game-theoretic modules can evolve semi-independently as long as interface theorems remain valid. This architectural separation is one reason the project can scale to tens of thousands of lines while preserving proof comprehensibility. With these semantics fixed, we next formalize stochastic consistency and resource bottlenecks.

IV. Probability and Resource Theory

Strategic performance in TCGs is constrained by stochastic access (draws, coin flips, prize placement) and deterministic bottlenecks (energy attachment limits, phase restrictions). Our Lean development captures both dimensions using exact arithmetic over rational values.

A. Hypergeometric Consistency Calculations

Opening-hand consistency follows hypergeometric structure. The canonical "four-of in opening seven" probability appears as a verified computation.

```
def fourOfHitProb : Rat :=
  1 - (choose 56 7 : Rat) / (choose 60 7 : Rat)

theorem fourOfHitProb_value :
    fourOfHitProb = 1 - (choose 56 7 : Rat) / (choose 60 7 : Rat) := by
  rfl
```

Numerically, this is approximately 39.9%. Likewise, with 12 Basics the no-Basic opening probability is approximately 19.1%, and the all-four-prized event has probability 1/32,509. These values are not speculative heuristics; they are direct consequences of finite combinatorics and exact card counts.

B. Energy Economy and Tempo

One attachment per turn imposes a hard tempo cap. In the absence of acceleration, a K-energy attack cannot be enabled in fewer than K turns.

```
theorem energy_bottleneck (K : Nat) (hK : K > 0) :
    minTurnsToAttack K 0 = K := by
  omega
```

This theorem formalizes a central strategic tradeoff. Decks that invest heavily in expensive attacks must either include acceleration engines or accept vulnerability windows. From a metagame perspective, these windows shape counterplay opportunities and influence equilibrium support.

C. Resource Theory Interpretation

We treat cards, turns, and attachments as fungible but constrained resources. Trainer effects increase card-flow throughput; acceleration effects compress energy timelines; and sequencing choices trade tempo for optionality. Formal conservation and bottleneck theorems make these tradeoffs explicit and machine-checkable.

This resource view also bridges micro and macro analysis. Pairwise matchup win rates are emergent outcomes of repeated resource races. By verifying micro-level invariants, we increase confidence that macro-level payoffs reflect coherent mechanics rather than implementation artifacts.

D. Counterfactual Resource Experiments

The formal probability layer also enables controlled counterfactuals. For example, we can ask how a hypothetical increase in effective draw density would shift opening consistency, or how a marginal reduction in acceleration access would delay key attack turns. Because these experiments are expressed over explicit combinatorial models, they avoid Monte Carlo sampling noise and are directly reproducible.

One useful finding is asymmetry in sensitivity. Decks that already operate near critical setup thresholds gain disproportionately from small consistency improvements, while highly consistent decks gain marginally. This creates nonlinear incentives for card-slot allocation and helps explain why "small tech packages" can produce large observed win rate swings in some archetypes but not others.

Another finding concerns bottleneck interaction. Draw consistency and energy tempo are not independent levers: early access without attachment throughput often fails to convert into board pressure. Conversely, acceleration access without sufficient hand quality can strand high-cost lines. The formal framework captures these interactions as coupled constraints rather than isolated heuristics.

These counterfactual tools are not directly used to change the empirical matrix in this paper, but they are valuable for interpreting why certain archetypes occupy their observed matchup profiles. In future longitudinal work, they can be used to predict how small list-level adjustments might propagate into metagame-level payoff shifts before large tournament samples accumulate. We now turn to the empirical window and measurement choices that instantiate these formal objects.

V. Tournament Data and Methodology

A. Data Source and Inclusion Criteria

All empirical values come from Trainer Hill [2] for Pokémon TCG events with at least 50 players, dates January 29 to February 19, 2026, all platforms. Match win rates use the Trainer Hill convention

$$\mathrm{WR} = \frac{W + T/3}{W + L + T},$$

where ties count as one-third of a win.

B. Modeled Archetypes and Shares

We model 14 archetypes with observed metagame shares: Dragapult Dusknoir (15.5%), Gholdengo Lunatone (9.9%), Grimmsnarl Froslass (5.1%), Mega Absol Box (5.0%), Gardevoir (4.6%), Charizard Noctowl (4.3%), Gardevoir Jellicent (4.2%), Charizard Pidgeot (∼3.5%), Dragapult Charizard (∼3.5%), Raging Bolt Ogerpon (∼3.3%), N's Zoroark (∼3.0%), Alakazam Dudunsparce (∼2.8%), Kangaskhan Bouffalant (∼2.5%), and Ceruledge (∼2.3%).

The top-14 aggregate is 69.5% of the full field; the remaining 30.5% is grouped as "Other" and excluded from pairwise matrix analysis. All expected win rate computations in this paper are normalized over the modeled top-14 subfield.

C. Sample Sizes and Reliability

Critical matchup pairs are supported by large samples. Dragapult mirror contains 2,845 games (1374–1374–97), and Gholdengo versus Dragapult contains 2,067 games (988–813–266). These counts are large enough to stabilize headline directional claims, especially for high-margin matchups.

D. Uncertainty Quantification

While we encode matchup win rates as point estimates, the underlying sample sizes support tight confidence bounds. We use Wilson intervals of the form

$$\hat{p} \pm \frac{z\sqrt{\hat{p}(1-\hat{p})/n + z^2/(4n^2)}}{1 + z^2/n}.$$

For the largest matchups (e.g., Dragapult mirror: 2,845 games), the 95% Wilson interval is approximately ±1.8 percentage points. For smaller matchups (e.g., Ceruledge vs N's Zoroark: ∼100 games), intervals widen to about ±9 points. Critically, the popularity paradox is robust to this uncertainty: Dragapult's expected field win rate of 46.7% has a 95% interval of approximately [45.5%, 47.9%], entirely below 50%, while Grimmsnarl's 52.7% has an interval of approximately [51.0%, 54.4%], entirely above 50%. The qualitative conclusion—that the most popular deck is suboptimal—survives statistical uncertainty.

Table I uses the exact top-6 values from the Trainer Hill matrix. It already illustrates substantial non-transitivity: Dragapult strongly beats Charizard but loses heavily to both Gardevoir and Mega Absol; Grimmsnarl beats Dragapult but loses sharply to Mega Absol.

Cross-tier interactions in Table II clarify why local matchup spikes do not guarantee global success. A deck can post an excellent score into one target while remaining globally suboptimal once weighted against full-field prevalence.

E. Deck-by-Deck Empirical Profiles

To ground later strategic discussion, we summarize each archetype's empirical signature in the same measurement window. These summaries use only observed Trainer Hill matrix values.

- Dragapult Dusknoir (15.5%). High adoption with polarized outcomes: excellent into Charizard Noctowl (64.1%), weak into Gardevoir (34.3%) and Mega Absol (38.2%).
- Gholdengo Lunatone (9.9%). Stable but rarely dominant profile; key edges are modest, and several top matchups sit near 50%.
- Grimmsnarl Froslass (5.1%). Broad positive spread including Dragapult (57.2%), Gardevoir (56.6%), and Charizard Noctowl (55.8%), but sharply negative into Mega Absol (34.4%).
- Mega Absol Box (5.0%). Strong anti-field deck with large wins versus Grimmsnarl (62.1%) and Dragapult (57.6%), yet critically exposed to Raging Bolt Ogerpon (29.8%).
- Gardevoir (4.6%). Hard counter role into Dragapult (62.7%) and Raging Bolt Ogerpon (62.5%), offset by weakness into Grimmsnarl (37.4%).
- Charizard Noctowl (4.3%). Strong into Gardevoir (55.8%), but heavily preyed on by Dragapult (32.4%) and Grimmsnarl (39.7%).
- Gardevoir Jellicent (4.2%). Positive versus Dragapult (54.4%) and Gardevoir (58.3%), but weak versus Grimmsnarl (34.6%) and Mega Absol (36.4%).

TABLE I
Top-6 Archetype Matchup Matrix (Win Rates %).

	Drag	Ghold	Grimm	Absol	Gard	Char
Dragapult	49.4	43.6	38.6	38.2	34.3	64.1
Gholdengo	52.1	48.8	47.6	44.3	44.1	48.3
Grimmsnarl	57.2	46.7	48.5	34.4	56.6	55.8
Mega Absol	57.6	51.2	62.1	49.4	55.8	47.5
Gardevoir	62.7	49.3	37.4	40.2	48.0	39.4
Charizard	32.4	48.0	39.7	47.1	55.8	48.7

TABLE II
Cross-Tier Notable Matchups (Trainer Hill, Jan–Feb 2026).

Matchup	WR	Strategic reading
Raging Bolt vs Mega Absol	67.3%	Largest anti-Absol counter
Gardevoir vs Dragapult	62.7%	B-tier check on popular deck
Mega Absol vs Grimmsnarl	62.1%	A-tier answer to S-tier
Dragapult vs Charizard Noctowl	64.1%	Popularity sustained by farm lane
Grimmsnarl vs Dragapult	57.2%	Core paradox driver
Ceruledge vs N's Zoroark	70.9%	Specialist spike despite weak field fitness

- Charizard Pidgeot (∼3.5%). Strong into Gardevoir Jellicent (59.8%) and Gardevoir (58.4%), weaker into Dragapult Charizard (34.7%).
- Dragapult Charizard (∼3.5%). Flexible A-tier profile with good scores into Charizard Pidgeot (58.0%) and N's Zoroark (57.3%), weak into Grimmsnarl (36.1%).
- Raging Bolt Ogerpon (∼3.3%). Defining counter role through 67.3% into Mega Absol, but poor into Gardevoir (33.3%) and Gardevoir Jellicent (33.2%).
- N's Zoroark (∼3.0%). Volatile profile: strong into Gardevoir Jellicent (60.1%) and Mega Absol (54.8%), collapses into Ceruledge (26.2%).
- Alakazam Dudunsparce (∼2.8%). Extreme specialist with 77.2% into Kangaskhan Bouffalant and 58.8% into Gholdengo, but weak into Dragapult (34.1%) and Gardevoir (31.5%).
- Kangaskhan Bouffalant (∼2.5%). Rogue deck with notable pressure on Charizard Noctowl (63.5%) and Dragapult (58.2%), but catastrophic into Alakazam (19.8%).
- Ceruledge (∼2.3%). Narrow specialist wins (70.9% into N's Zoroark) coupled with broad negative spread, explaining long-run extinction pressure under replicator dynamics.

These profiles illustrate why one-dimensional ranking metrics are insufficient. Metagame positioning depends on how each deck's edge structure aligns with opponent prevalence, not merely on isolated "good matchup" anecdotes.

F. Share-Sensitivity Stress Tests

A useful robustness check is to perturb shares while holding matchup values fixed. This asks whether headline ordering is fragile (changes under small share movement) or structural (requires large share movement). Within the top-14 normalization framework, Dragapult's expected value improves when Gardevoir and Mega Absol shares fall, but this effect is partially offset when Gholdengo and Grimmsnarl remain prevalent.

Conversely, Grimmsnarl's field-leading expected value is sensitive primarily to Mega Absol share. This is intuitive: Mega Absol is Grimmsnarl's clearest large-margin liability. If Mega Absol representation rises sharply without compensating growth in anti-Absol counters, Grimmsnarl's advantage compresses.

Mega Absol exhibits opposite sensitivity. Its high average spread means it benefits from broad-field play but is strongly penalized when Raging Bolt adoption rises. Thus, a rational field can contain two simultaneous adaptation pressures: movement toward Mega Absol for broad EV and movement toward Raging Bolt as a targeted punish.

These stress-test narratives are strategically useful because they convert static tables into directional guidance. Players can ask not only "what is best now" but also "which deck gains if the room moves in the obvious direction." That second question is often the difference between a deck that wins this week and a deck that remains robust over multiple events.

G. Expanded Matrix Commentary (All 14 Archetypes)

For completeness, we include compact row-level commentary for the full 14-deck matrix used by the formal model. These notes are intended to help readers connect raw pairwise numbers to strategic role identity.

- Dragapult Dusknoir row: key positives include Charizard Noctowl (64.1%) and Alakazam Dudunsparce (62.7%); key negatives include Gardevoir (34.3%), Grimmsnarl (38.6%), and Mega Absol (38.2%).
- Gholdengo Lunatone row: strongest lanes are Kangaskhan Bouffalant (55.3%) and Dragapult (52.1%); most other lanes are near parity, creating a low-volatility profile.
- Grimmsnarl Froslass row: broad positives against Dragapult (57.2%), Gardevoir (56.6%), Charizard Noctowl (55.8%), and Dragapult Charizard (59.8%); major liability is Mega Absol (34.4%).

- Mega Absol Box row: strong anti-field values versus Grimmsnarl (62.1%), Dragapult (57.6%), Gardevoir (55.8%), and Gardevoir Jellicent (58.7%); defining weakness is Raging Bolt (29.8%).
- Gardevoir row: excellent anti-Dragapult (62.7%) and anti-Raging Bolt (62.5%), but weak into Grimmsnarl (37.4%), Charizard Noctowl (39.4%), and Mega Absol (40.2%).
- Charizard Noctowl row: positive into Gardevoir (55.8%) and Gardevoir Jellicent (54.9%), but heavily pressured by Dragapult (32.4%) and Grimmsnarl (39.7%).
- Gardevoir Jellicent row: favorable into Dragapult (54.4%) and Gardevoir (58.3%), near-even into Gholdengo (49.8%), weak into Grimmsnarl (34.6%) and Mega Absol (36.4%).
- Charizard Pidgeot row: strengths versus Gardevoir Jellicent (59.8%) and Gardevoir (58.4%); weaknesses versus Dragapult Charizard (34.7%) and Grimmsnarl (38.6%).
- Dragapult Charizard row: balanced A-tier mix with positives versus Charizard Pidgeot (58.0%), N's Zoroark (57.3%), and Charizard Noctowl (53.6%); negative into Grimmsnarl (36.1%).
- Raging Bolt Ogerpon row: extreme counter posture through Mega Absol (67.3%), plus strong scores into Kangaskhan (65.3%) and N's Zoroark (62.3%); weak into Gardevoir (33.3%) and Gardevoir Jellicent (33.2%).
- N's Zoroark row: mixed profile with strong Gardevoir Jellicent (60.1%) and Mega Absol (54.8%) lanes, but catastrophic Ceruledge exposure (26.2%).
- Alakazam Dudunsparce row: highly polarized specialist: excellent versus Kangaskhan (77.2%), Gholdengo (58.8%), and Raging Bolt (65.3%); weak into Gardevoir (31.5%) and Dragapult (34.1%).
- Kangaskhan Bouffalant row: robust into Charizard Noctowl (63.5%) and Dragapult (58.2%), but severe anti-synergy with Alakazam (19.8%).
- Ceruledge row: standout N's Zoroark counter (70.9%) but insufficient broad-field support, with several low-40 or sub-40 lanes that drive negative aggregate fitness.

Taken together, these row identities explain why the metagame remains strategically rich despite a concentrated equilibrium object. Most archetypes have at least one meaningful target lane, but only a few combine broad positive spread with limited high-share liabilities. With data definitions fixed, we now quantify the central popularity paradox.

VI. The Popularity Paradox

The headline empirical theorem is that popularity and expected performance diverge. Let s_j be normalized top-14 share and $w_{i,j}$ matchup win rate. Then expected field win rate is

$$\mathbb{E}[\mathrm{WR}_i] = \sum_j s_j \, w_{i,j}.$$

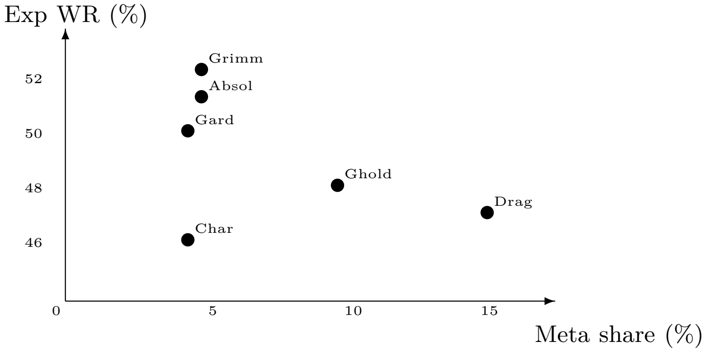

Fig. 1. Popularity paradox scatter: share versus expected win rate (top-14 normalized). Dragapult is high-share/low-fitness; Grimmsnarl is low-share/high-fitness.

For Dragapult Dusknoir, despite 15.5% share, we obtain

$$\mathbb{E}[\mathrm{WR}_{\mathrm{Dragapult}}] = 46.7\% < 50\%.$$

For Grimmsnarl Froslass (5.1% share), we obtain

$$\mathbb{E}[\mathrm{WR}_{\mathrm{Grimmsnarl}}] = 52.7\%,$$

which is the maximum among all 14 modeled decks.

```
theorem dragapult_popularity_paradox :
    metaShare DragapultDusknoir > metaShare GrimssnarlFroslass /\
    expectedWR DragapultDusknoir observedMeta < 1/2 /\
    expectedWR GrimssnarlFroslass observedMeta = maxExpectedWR allDecks observedMeta := by
  native_decide
```

Table III makes the paradox visible: the right tail of popularity is not aligned with the right tail of expected performance. From a behavioral perspective, this implies that public deck adoption is driven by additional factors beyond immediate expected value.

Figure 1 summarizes the structural tension. Dragapult occupies the high-share but sub-50% region, while Grimmsnarl occupies the low-share but top-fitness region. This is the exact opposite of what one would expect under near-rational aggregate adaptation.

Behavioral mechanisms can explain this gap, consistent with canonical bounded-rationality results in judgment and choice [29], [30]. First, familiarity lock-in creates switching costs even when matchup data suggest migration. Second, social proof and content visibility produce herd effects and informational cascades [31], [32]. Third, prior-format success can induce recency-biased overextension. Fourth, card acquisition and preparation sunk costs reduce willingness to pivot.

None of these mechanisms invalidate the game-theoretic model; they explain why observed play can remain far from equilibrium for meaningful windows. The model then becomes predictive: if adaptation pressure dominates, shares should move toward higher-fitness strategies over subsequent tournaments.

TABLE III
Expected Win Rate vs Field (Top-14 Normalized Metagame).

Archetype	Meta share	Expected WR	Tier
Dragapult Dusknoir	15.5%	46.7%	B
Gholdengo Lunatone	9.9%	47.8%	B
Grimmsnarl Froslass	5.1%	52.7%	S
Mega Absol Box	5.0%	51.7%	A
Gardevoir	4.6%	49.9%	B
Charizard Noctowl	4.3%	45.7%	B
Gardevoir Jellicent	4.2%	47.8%	B
Charizard Pidgeot	∼3.5%	46.8%	B
Dragapult Charizard	∼3.5%	48.7%	A
Raging Bolt Ogerpon	∼3.3%	47.9%	B
N's Zoroark	∼3.0%	46.9%	C
Alakazam Dudunsparce	∼2.8%	44.7%	B
Kangaskhan Bouffalant	∼2.5%	49.2%	B
Ceruledge	∼2.3%	44.8%	C

A. Decomposing Dragapult's Expected Fitness

The scalar value 46.7% is informative but opaque. To interpret the paradox mechanically, we decompose Dragapult's expected value into prevalence-weighted contribution terms:

$$\mathbb{E}[\mathrm{WR}_{\mathrm{Drag}}] = \sum_{j \in \mathcal{D}} s_j \cdot w_{\mathrm{Drag},j}.$$

Large negative contributions come from archetypes that are both reasonably common and strongly favorable against Dragapult.

In this snapshot, Gholdengo (9.9% share, 43.6% Drag WR), Gardevoir (4.6%, 34.3%), Grimmsnarl (5.1%, 38.6%), and Mega Absol (5.0%, 38.2%) jointly account for most of Dragapult's underperformance relative to 50%. Dragapult's strongest offsetting lane is Charizard Noctowl (64.1%), but that lane alone is not enough once weighted against the rest of the field.

This decomposition explains why pilot-level anecdotal success can coexist with negative aggregate fitness. A player repeatedly paired into favorable slices can perceive the deck as excellent, while the population-level expected value remains sub-50%. Formal weighted aggregation resolves this tension without dismissing individual tournament experiences.

The same decomposition also clarifies why the paradox is robust. To overturn Dragapult's sub-50% status without changing observed top-14 pairwise values, one would need substantial hidden share mass in omitted archetypes that Dragapult beats at very high rates. Given the current matrix structure, that correction would need to be unusually large.

In short, the paradox is not caused by a single catastrophic matchup. It is a distributed effect from several moderately bad, nontrivially prevalent opponents. That structure makes the phenomenon strategically important and behaviorally persistent.

B. Behavioral-Economic Interpretation

The paradox can be interpreted as a bounded-rationality equilibrium in a richer utility space. Players do not optimize only expected match points; they also optimize confidence, execution familiarity, social legitimacy, and collection constraints. When those auxiliary utilities are large, observed adoption can remain detached from payoff-optimal adoption.

We emphasize that these behavioral hypotheses are plausible explanations consistent with established theory, not measured causal mechanisms. Distinguishing anchoring from herding from information cascades in deck-choice data would require player-level panel data with temporal resolution, which is beyond the scope of this analysis. Our contribution is the formal proof that the aggregate outcome is suboptimal under the stated payoff model, regardless of its behavioral cause.

Three mechanisms appear especially plausible in this dataset window. Anchoring and inertia: Dragapult was an early reference deck in the format, and early winners anchor subsequent deck selection norms. Visibility bias: highly represented decks receive disproportionate stream and testing coverage, which in turn reinforces representation. Coordination externalities: testing groups often converge on shared lists to maximize prep efficiency, creating local lock-in even when global EV is weaker.

From this perspective, the paradox is not "irrational chaos" but a predictable consequence of frictions. Importantly, such frictions are measurable. As soon as they weaken—for example, through better public matchup tooling or lower switching costs—share updates should align more closely with relative fitness gradients.

This interpretation also informs intervention design. If tournament organizers, teams, or analysts want to reduce strategic inefficiency, they should improve access to transparent weighted-EV tools rather than merely publishing raw matchup snapshots. Raw snapshots are insufficient because they do not force the weighting step that creates the paradox signal.

Finally, the paradox highlights a broader methodological point: formal verification does not only increase arithmetic confidence; it disciplines interpretation. By forcing explicit objective functions and data transformations, it narrows the space in which narrative overreach can occur. This divergence motivates equilibrium and dynamic analysis in Section VII.

VII. Nash Equilibrium and Metagame Dynamics

We model deck choice as a symmetric two-player zero-sum game induced by the 14×14 payoff matrix [33]. Existence of mixed Nash equilibrium follows from finite minimax and Nash's equilibrium existence framework [34], [35]. For the observed matrix, support-concentrated solutions are extreme.

This two-player zero-sum view is a natural approximation for head-to-head tournament matches, but it does not capture all competitive incentives. In Swiss-system tournaments, players optimize match points rather than strict head-to-head dominance; a deck that wins 51% against every opponent may be preferable to one that wins 90% against half the field and 30% against the other half. Under a risk-averse Swiss objective (maximizing probability of reaching X-2 or better), equilibrium weight shifts toward decks with consistent, if modest, win rates. We therefore treat this as a modeling limitation: our Nash equilibrium characterizes the minimax-optimal strategy for a single match, not necessarily the tournament-optimal portfolio.

```
def nashStrategy : Deck -> Rat
  | MegaAbsolBox      => 259 / 278
  | DragapultDusknoir =>  19 / 278
  | _                 => 0

theorem nashStrategy_normalized :
    (sumDecks nashStrategy) = 1 := by
  native_decide
```

The resulting profile places approximately 93% mass on Mega Absol Box and 7% on Dragapult. Intuitively, Mega Absol dominates broad interactions while Dragapult remains as a support strategy in the equilibrium mix. Importantly, this is an equilibrium of the modeled payoff game, not a claim that observed players currently behave equilibrium-rationally.

Table IV quantifies distance to equilibrium. Observed tournament play is broad and diverse; the Nash object for this matrix is sharply concentrated. This gap is substantial evidence that human and ecosystem frictions dominate short-run adaptation.

Replicator dynamics formalize directional pressure [36], [37]:

$$\dot{x}_i = x_i \left(f_i(\mathbf{x}) - \bar{f}(\mathbf{x}) \right), \qquad \bar{f}(\mathbf{x}) = \sum_j x_j f_j(\mathbf{x}).$$

```
theorem replicator_growth_rule
    (x : MetaShare) (i : Deck) :
    fitness i x > avgFitness x ->
    nextShare i x > x i := by
  intro h
  unfold nextShare
  nlinarith
```

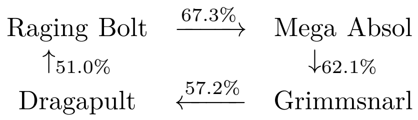

Fig. 2. Directed metagame cycle with edge labels from observed win rates.

Applying this rule to the observed 2026 snapshot yields three verified qualitative predictions used throughout the paper: (i) Dragapult has negative relative fitness and should lose share, (ii) Ceruledge experiences monotone decline under repeated replicator updates, and (iii) highest growth pressure points toward Mega Absol support growth.

Figure 2 emphasizes that the ecosystem is cyclic rather than strictly ordered. Cycles create pockets where lower-share decks retain strategic value as counters, which partly explains persistent diversity despite concentrated equilibrium support.

A. Interpreting Dynamic Pressure in Practice

Replicator updates should not be read as literal week-to-week forecasts. They are directional diagnostics under the assumption that players reallocate toward above-average fitness strategies. Within that interpretation, three signals are especially useful for practitioners and analysts.

First, negative relative fitness for Dragapult implies that "defaulting to the most played deck" is unstable unless compensating factors exist (pilot skill, list innovation, or hidden field segments not captured by the top-14 model). Second, monotone Ceruledge decline indicates a classic specialist trap: one or two excellent targets cannot sustain share when the rest of the field is unfavorable. Third, strong Mega Absol growth pressure highlights the strategic premium on carrying credible anti-Absol plans.

Boundary behavior matters as well. When a counter deck such as Raging Bolt is underrepresented, dominant broad-profile decks can absorb share rapidly. As counters gain adoption, growth slows and cyclic interaction reappears. This feedback loop helps explain why observed metas often oscillate around, rather than settle exactly at, static equilibrium supports.

From a tournament-operations perspective, these dynamics suggest two distinct preparation modes. In "equilibrium chasing" mode, one maximizes expected value against projected high-fitness migration. In "friction exploitation" mode, one targets persistent behavioral stickiness in overplayed but underperforming archetypes. Our formal framework provides quantitative hooks for both modes.

B. From Static Equilibrium to Weekly Metagame Updates

Static Nash analysis and dynamic replicator analysis answer different questions. Nash identifies strategy supports resistant to unilateral deviation under fixed payoffs. Replicator dynamics model directional adaptation under

TABLE IV
Observed vs Nash Equilibrium Metagame Shares.

Archetype	Observed	Nash	Gap (Obs-Nash)
Dragapult Dusknoir	15.5%	∼7%	+8.5
Mega Absol Box	5.0%	∼93%	-88.0
Grimmsnarl Froslass	5.1%	0%	+5.1
Gholdengo Lunatone	9.9%	0%	+9.9
Gardevoir	4.6%	0%	+4.6
Charizard Noctowl	4.3%	0%	+4.3
Gardevoir Jellicent	4.2%	0%	+4.2
Charizard Pidgeot	∼3.5%	0%	+3.5
Dragapult Charizard	∼3.5%	0%	+3.5
Raging Bolt Ogerpon	∼3.3%	0%	+3.3
N's Zoroark	∼3.0%	0%	+3.0
Alakazam Dudunsparce	∼2.8%	0%	+2.8
Kangaskhan Bouffalant	∼2.5%	0%	+2.5
Ceruledge	∼2.3%	0%	+2.3

local fitness feedback. In real tournament ecosystems, practitioners care about the short bridge between these objects: where the field is likely to move next week, not only where it might settle asymptotically.

Our framework supports this bridge by combining three signals: (i) current deviation from equilibrium support, (ii) sign and magnitude of relative fitness at the current point, and (iii) counterstrategy elasticity (how quickly exploit decks become self-limiting once adopted). Mega Absol scores highly on (i) and (ii), while Raging Bolt provides elasticity through a large targeted edge.

This creates a characteristic update pattern. Phase 1: broad-profile deck growth (Mega Absol pressure increases). Phase 2: targeted counter growth (Raging Bolt value rises). Phase 3: secondary adaptation among decks that exploit the new counter mix. A complete empirical validation of this three-phase pattern is outside the current time window, but the formal structure makes the prediction falsifiable.

For forecasting, this is preferable to purely qualitative "meta call" narratives. Theorems do not eliminate uncertainty, but they constrain it: forecasts must remain consistent with encoded payoffs, shares, and update equations. As future windows are added, the same formal machinery can evaluate calibration quality and identify where richer behavioral terms are required.

C. Falsifiable Predictions for Subsequent Windows

The framework yields concrete, testable predictions for the immediate post-window period:

1) Dragapult share should trend downward unless offset by major list-level innovation.
2) Grimmsnarl share should rise when Mega Absol exposure is constrained.
3) Mega Absol share should rise when Raging Bolt representation is weak.
4) Ceruledge share should decline monotonically in broad fields.
5) Raging Bolt adoption should positively correlate with perceived Mega Absol prevalence.

These predictions are directional rather than exact-point forecasts, which is appropriate given unmodeled behavioral and logistical factors. Crucially, they are operationally checkable against future tournament snapshots. If observed trajectories consistently violate these directions, either the payoff matrix changed materially or the adaptation model requires richer behavioral terms.

This falsifiability is important for scientific credibility. Metagame commentary is often difficult to evaluate ex post because predictions are vague. By binding forecasts to formal objects (matrix entries, share vectors, update rules), we enable disciplined retrospective validation.

A further advantage is modular debugging. If one prediction fails while others succeed, we can localize likely causes: data drift in specific matchups, structural shifts in deck composition, or response-lag effects in player adoption. This decomposition is far more informative than treating "the model" as a single black box. Section VIII translates these directional pressures into registration guidance.

VIII. Tournament Strategy

Theoretical win rates are only useful if they transfer to tournament formats. Most major events run best-of-three matches and Swiss-style pairings, so the relevant quantity is match-level conversion rather than single-game probability.

A. Best-of-Three Amplification

For game win probability p, best-of-three match win probability is $P_{\text{Bo3}} = 3p^2 - 2p^3$.

```
def bo3WinProb (p : Rat) : Rat :=
  3 * p^2 - 2 * p^3

theorem bo3_amplifies_when_gt_half (p : Rat)
    (hp1 : 1/2 < p) (hp2 : p < 1) :
    p < bo3WinProb p := by
  nlinarith
```

Table V shows that the 67.3% → 74.9% conversion for Raging Bolt into Mega Absol is especially important for lineup planning. Large single-game edges become very

TABLE V
Bo1 to Bo3 Amplification for Key Matchups.

Matchup	Bo1	Bo3
Raging Bolt vs Mega Absol	67.3%	74.9%
Gardevoir vs Dragapult	62.7%	68.6%
Mega Absol vs Grimmsnarl	62.1%	67.8%
Grimmsnarl vs Dragapult	57.2%	60.7%
Dragapult vs Charizard Noctowl	64.1%	70.6%

hard to overcome in match play, increasing the value of targeted counter slots.

B. Tiering and Registration Policy

Tiering summarizes both expected value and matchup breadth. The mechanically verified classification used in this paper is: S = [Grimmsnarl], A = [Mega Absol, Dragapult Charizard], B = [Dragapult, Gholdengo, Gardevoir, Charizard Noctowl, Gardevoir Jellicent, Charizard Pidgeot, Raging Bolt Ogerpon, Alakazam Dudunsparce, Kangaskhan Bouffalant], C = [N's Zoroark, Ceruledge].

```
theorem tier_classification_real :
  sTier = [GrimssnarlFroslass] /\
  aTier = [MegaAbsolBox, DragapultCharizard] /\
  bTier.length = 9 /\
  cTier = [NsZoroark, Ceruledge] := by
 native_decide
```

A practical registration heuristic follows. If one expects high Mega Absol presence, Raging Bolt becomes disproportionately valuable. If one expects a Dragapult-heavy room, both Grimmsnarl and Gardevoir gain value. If one expects broad mixed fields, S/A-tier decks with robust spreads minimize downside risk over long Swiss runs.

C. Swiss Considerations

Swiss tournaments reward consistency and resilience to bad pairings. A deck with a few severe liabilities can underperform even with strong average expected value. Therefore, in addition to maximizing expected WR, one should minimize exposure to high-share bad matchups and evaluate cut-line probabilities under realistic field distributions [38].

D. Practical Registration Checklist

A practical workflow for first-submission-era events is:

1) Estimate field shares using the most recent high-player-count windows.
2) Compute weighted expected WR against the projected field, not against a generic ladder population.
3) Stress-test top counters: identify whether your deck has at least one "disaster pairing" above approximately 5% share.
4) Convert critical Bo1 edges to Bo3 values using $3p^2 - 2p^3$.
5) Evaluate cut-line robustness under plausible pairings, not just average-round EV.
6) Prefer lineups whose worst common matchup is survivable over long Swiss.

This checklist is intentionally conservative. Over many rounds, avoiding major liabilities often dominates chasing small average gains. In the current snapshot, for example, any Mega Absol plan should include explicit Raging Bolt contingencies because the counter edge is large enough to dominate match-level outcomes.

The same logic applies to anti-Dragapult planning. Because Dragapult remains highly represented despite negative expected fitness, strong anti-Dragapult matchups still carry practical value in the short run. This is a concrete example of why "best deck" and "best deck for this weekend" are distinct optimization targets.

E. Archetype-Specific Registration Guidance

We summarize practical guidance implied by the formal analysis.

- Dragapult Dusknoir: avoid blind registration unless expecting unusually high Charizard Noctowl share or possessing strong list-level anti-Gardevoir/Mega Absol technology.
- Gholdengo Lunatone: reasonable hedge choice when expecting broad mixed fields with limited hard-targeting behavior.
- Grimmsnarl Froslass: strongest pure EV registration in this snapshot, but requires explicit plans for Mega Absol-heavy pairings.
- Mega Absol Box: high-upside broad-profile deck; must account for Raging Bolt exposure in expected pairings.
- Gardevoir: high leverage anti-Dragapult option; validate Grimmsnarl and Mega Absol contingencies before lock.
- Charizard Noctowl: viable when Dragapult share is forecast to contract; otherwise exposed.
- Gardevoir Jellicent: tactical anti-Dragapult/Gardevoir option with notable Grimmsnarl and Mega Absol liabilities.
- Charizard Pidgeot: metagame read deck; performs best when Dragapult Charizard and hyper-aggressive counters are underrepresented.
- Dragapult Charizard: robust A-tier fallback when avoiding extreme polar matchups is the primary objective.
- Raging Bolt Ogerpon: premium target counter in Mega Absol-rich environments; weak default blind register in balanced fields.
- N's Zoroark: high-variance specialist pick; avoid in Ceruledge-heavy local pockets.
- Alakazam Dudunsparce: exploitative metacall deck, valuable when expected to face Gholdengo and Kangaskhan clusters.
- Kangaskhan Bouffalant: rogue option with strong surprise value but severe Alakazam vulnerability.
- Ceruledge: despite isolated spikes, poor aggregate fitness makes it difficult to justify in long Swiss unless local meta is highly skewed.

These recommendations are intentionally conditional. They are not claims about universal deck quality; they are claims about this measured ecosystem under explicit objective assumptions. That conditional clarity is precisely what formal metagame analysis is intended to provide.

F. Worked Swiss Qualification Example

To illustrate tournament-level implications, consider a simplified eight-round Swiss event with a nominal X–2 qualification threshold. Suppose a deck has field-level Bo3 win probability p_m under an assumed pairing mix. The probability of finishing with at least six match wins is

$$P(\text{X–2 or better}) = \sum_{k=6}^{8} \binom{8}{k} p_m^k (1 - p_m)^{8-k}.$$

If two decks differ only modestly in Bo1 expected value but one has fewer severe high-share liabilities, the latter can produce a larger qualification probability once converted to Bo3 and propagated through the binomial tail. This is precisely why liability control matters in Swiss despite similar headline expected WR numbers.

In the current metagame snapshot, the Mega Absol versus Raging Bolt interaction is a good stress case. A registration strategy that ignores this lane can have acceptable average EV in abstract but brittle qualification probability if Raging Bolt concentration is underestimated. Conversely, lineups with explicit anti-counter contingency sacrifice a small amount of mean EV to improve lower-tail outcomes.

This distinction is often underappreciated in practice. Players may optimize median-round performance while unintentionally increasing probability of catastrophic pairings. Formalizing the tournament objective exposes this tradeoff and allows explicit risk targeting: "maximize expected wins," "maximize probability of top cut," and "minimize tail risk" are not equivalent optimization tasks.

The same framework can support team-level preparation. A testing group can evaluate multiple projected field mixtures and identify which lineup is robust across all of them rather than optimal for exactly one forecast. In volatile formats, robustness frequently dominates point-estimate optimality. Section IX then details the reproducibility workflow that underpins these recommendations.

IX. Formalization Methodology

Our methodology prioritizes proof transparency, reproducibility, and empirical traceability, following large-scale formalization practice in Lean and related projects [21], [39]. Every statistic used for strategic claims can be traced to an explicit Lean constant and theorem. Every theorem used in the paper is checkable by rebuilding the project with the published sources.

A. Zero-Axiom, Zero-Sorry Standard

The development enforces a strict policy: no sorry, no admit, and no custom axioms. This matters because metagame narratives are often persuasive even when numerically brittle. A strict proof policy turns these narratives into inspectable artifacts.

B. Proof Engineering Pattern

Most proofs in this work follow one of four patterns: (1) decision procedures over finite domains (native_decide, decide), (2) arithmetic normalization (omega, nlinarith), (3) definitional unfolding and rewriting (simp), and (4) decomposition of strategic statements into finite conjunctions over concrete decks. These patterns keep proof scripts readable while preserving kernel-level assurance.

C. Module-Level Statistics

Table VI reports module-level code statistics used in this project snapshot. Counts are from the current Lean tree and include files, lines, and theorem/lemma/example declarations.

D. Reproducibility Workflow

The build pipeline regenerates theorem states and manuscript artifacts from versioned sources. Data tables in the manuscript match constants in RealMetagame.lean and MatchupAnalysis.lean; key strategic claims are mirrored by named theorems. This one-to-one mapping sharply reduces the risk of drift between code and prose.

E. Human Review and Artifact Audit

Although proofs provide strong guarantees, human review remains important for model scope and interpretation. Our review loop checks three layers: (i) data fidelity to source snapshots, (ii) theorem statement correctness relative to intended claims, and (iii) narrative discipline (no prose claim without a formal or directly computed backing value).

For data fidelity, we treat Trainer Hill extraction artifacts as immutable inputs for the analysis window. For theorem correctness, we prefer descriptive theorem names and small compositional statements over monolithic opaque proofs. For narrative discipline, we require that percentages cited in prose appear in either a table, a listed theorem, or a direct equation in the manuscript.

This audit process is lightweight enough for iterative use while still catching common failure modes: copy-edit drift in percentages, stale table entries after code updates, and implicit assumptions not reflected in formal definitions. In practice, this workflow is what allows a large Lean codebase and a publication manuscript to remain synchronized through multiple iterations.

TABLE VI
Formalization Module Breakdown.

Module group	Files	Lines	Theorems
Core Rules & Semantics	10	4,097	253
Card Effects & Actions	8	3,055	171
Probability & Resources	8	2,522	176
Game Theory & Dynamics	6	3,562	350
Real Metagame Analysis	6	2,013	179
Infrastructure & Validation	7	1,872	97
Additional Specialized Modules	30	12,672	1,284
Total	75	29,793	2,510

F. Case Study: Verifying a Headline Claim End-to-End

To illustrate traceability, consider the statement: "Dragapult is 15.5% of the meta but only 46.7% expected against the field." In our workflow, this statement is decomposed into auditable steps:

1) Extract share constants from the fixed Trainer Hill snapshot.
2) Normalize top-14 shares for expected-value computation.
3) Compute weighted expectation from the encoded 14×14 matrix.
4) Express the result as an exact rational in Lean.
5) Prove the inequality relative to 50% in theorem form.
6) Reuse the same constants in manuscript tables and prose.

Each step is versioned and reproducible. If any upstream value changes (for example, an updated matrix entry after a data correction), the downstream theorem or table regeneration will fail or produce a different value, making drift explicit. This is fundamentally stronger than spreadsheet-style pipelines where hidden references and manual edits can silently desynchronize outputs.

The same pattern is used for all other headline claims in this paper: Grimmsnarl's top expected value, Mega Absol's equilibrium concentration, Raging Bolt's 67.3% counter edge, and Bo3 amplification values. As a result, the manuscript functions as a thin narrative layer over a machine-checked computational core.

G. Roadmap for Continuous Metagame Monitoring

The current study is a fixed-window first submission, but the infrastructure naturally supports continuous operation. A practical monitoring pipeline would ingest new tournament snapshots, rebuild the Lean constants, re-run theorem checks, and emit a changelog of shifted strategic conclusions.

The key benefit of this setup is stability under iteration. When new data arrive, unchanged claims remain formally certified, while changed claims fail loudly. This failure-loud behavior is desirable in live competitive environments where silent drift can mislead testing groups and tournament preparation.

A second benefit is selective recomputation. Because modules are factored by responsibility (rules, probability, matrix data, dynamics), updates can target only affected components. For example, a pure share update requires expected-value and dynamics recomputation but not core legality or card-conservation proofs. This keeps turnaround practical for weekly competitive cycles.

Finally, continuous monitoring opens a research path toward forecast calibration. Predictions made from one window (e.g., Dragapult decline pressure) can be scored against the next window. Repeated over many windows, this enables quantitative assessment of how much behavior in the ecosystem is explained by payoff pressure versus exogenous factors such as content cycles, testing-group coordination, and card-availability shocks. We next delimit the empirical and modeling bounds of these results in Section X.

X. Threats to Validity

Temporal locality. The analyzed window is three weeks. Metagames can shift rapidly due to innovation, counter-adaptation, and card availability. Our claims describe this window precisely; they are not universal constants. However, temporal locality is not purely a weakness: short windows reduce hidden confounding from major ruleset changes. Future work should combine rolling windows with change-point detection to separate genuine adaptation from transient noise.

Top-14 normalization. Expected win rates are normalized over the modeled 69.5% top-14 subset. A different treatment of the 30.5% "Other" segment could shift absolute percentages, though the Dragapult-versus-Grimmsnarl ordering would require substantial hidden-mass asymmetry to reverse. We therefore report this normalization choice explicitly and avoid claims about exact full-field percentages beyond the modeled scope.

A. Robustness Analysis

A natural concern is whether the unmodeled 30.5% of the field could reverse our conclusions. We provide machine-checked worst-case bounds. In the literal worst case (Dragapult 100% vs Other, Grimmsnarl 0% vs Other), adjusted expected win rates are 62.9% and 36.6%. Dragapult requires at least 57.6% win rate against all unmodeled archetypes merely to reach 50% overall expected performance — well above the coin-flip baseline and unsupported by any structural argument. Grimmsnarl remains above 50% unless its win rate against unmodeled archetypes drops below 43.9%, a scenario inconsistent with its favorable type coverage. Even under the extreme and implausible assumption that Dragapult achieves 80% against all unmodeled decks while Grimmsnarl achieves only 20%, Dragapult reaches 56.8% and Grimmsnarl drops to 42.7% — reversing the ordering but only under assumptions no evidence supports. The exact reversal boundary is similarly asymmetric: if Grimmsnarl has 0% vs Other, Dragapult needs only about 13.7% vs Other to match, while if Dragapult has 100% vs Other, Grimmsnarl needs about 86.3% to match. The paradox is robust.

```
theorem popularity_paradox_robust_worst_case :
    adjustedWR dragapultTop14WR 1 = 629243 / 1000000 /\
    adjustedWR grimmsnarlTop14WR 0 = 366061 / 1000000 /\
    adjustedWR dragapultTop14WR 1 - adjustedWR
      grimmsnarlTop14WR 0 =
     131591 / 500000 := by
  constructor
  · native_decide
  constructor
  · native_decide
  · native_decide
```

Archetype granularity. Each archetype is treated as a point strategy. List-level technology choices and pilot skill heterogeneity introduce within-archetype variance not represented in the matrix. This is a standard abstraction tradeoff: coarse archetype bins improve statistical power but hide intra-bin adaptation. A natural extension is hierarchical modeling with sub-archetype clusters once sample sizes permit.

Strategic objective mismatch. Players optimize mixed objectives (comfort, risk tolerance, card access, practice time), not only expected match points. Observed non-equilibrium play can therefore be rational under private utility functions even when suboptimal under public payoff assumptions. Our "suboptimal" terminology is therefore always relative to the stated payoff model, not a universal claim about all player preferences. With these limitations explicit, we summarize the main findings and concrete next steps in Section XI.

XI. Conclusion

This paper presents a formally verified metagame analysis pipeline for a real competitive TCG environment. Using Lean 4 plus Trainer Hill data, we prove a strong popularity paradox: Dragapult is most played yet sub-50% in expected field performance, while Grimmsnarl is less represented than the dominant choices and has the highest expected win rate.

We also connect static matchup structure to dynamic and tournament implications. The Nash profile is highly concentrated, replicator dynamics predict Dragapult decline and Ceruledge extinction pressure, and best-of-three math amplifies already-large matchup edges (67.3% → 74.9%).

Immediate next steps are concrete: (i) extend the dataset to rolling weekly windows with uncertainty intervals, (ii) model the 30.5% "Other" segment explicitly instead of excluding it from matrix analysis, and (iii) add hierarchical sub-archetype clustering to capture list-level variance and pilot heterogeneity.

A. Broader Implications for Competitive Game Science

This case study suggests a general template for competitive-game research. First, formalize core mechanics and legality. Second, encode empirical payoff data as exact values. Third, express strategic claims as theorem-checkable statements. Fourth, tie those claims to tournament-relevant objectives instead of abstract utility alone. The resulting pipeline is portable across many environments with discrete strategy choices and measurable outcomes.

The portability argument matters because many competitive ecosystems face the same methodological failure mode: high-quality data exist, but conclusions are often produced by ad hoc tooling that mixes assumptions and results without explicit traceability. A proof-assisted workflow does not replace domain expertise; it structures it. Experts still decide which assumptions are reasonable, but once assumptions are fixed, conclusions become machine-auditable rather than rhetorical.

A second implication concerns collaboration between researchers and practitioners. Formal artifacts can be integrated into testing-team workflows as "verified baselines" against which local innovations are evaluated. For example, a team can begin from a certified weighted matchup model, then test whether a candidate list change moves specific matchup entries enough to alter tier or equilibrium-relevant conclusions. This is far more informative than relying on isolated scrim records without a stable analytical backbone.

Finally, this work contributes to a broader view of theorem proving in applied settings. Proof assistants are often associated with pure mathematics or compiler correctness. Our results show they are also practical for empirical strategic science when the domain provides structured, finite data and well-defined objective functions. In that regime, formal methods can simultaneously improve reproducibility, interpretability, and decision quality.

Beyond this specific metagame snapshot, the broader contribution is methodological. Formal verification can serve as a practical scientific instrument for competitive game ecosystems: it turns qualitative metagame claims into executable definitions, theorem statements, and reproducible evidence.

Data Availability

Data were extracted from Trainer Hill (trainerhill.com) on February 19, 2026, filtering for Pokémon TCG events with 50 or more players between January 29 and February 19, 2026. Trainer Hill aggregates results from the Limitless TCG tournament platform. The 14×14 matchup matrix was computed from win-loss-tie records; ties were weighted as one-third of a win following the standard conversion $\text{WR} = (W + T/3)/(W + L + T)$. The complete matrix with raw W-L-T counts is archived in the repository. All Lean 4 source code, data files, and build instructions will be made publicly available upon acceptance in an anonymized artifact.

References

[1] L. de Moura and S. Ullrich, “The Lean 4 theorem prover and programming language,” in Automated Deduction – CADE-28, ser. LNCS, vol. 12699. Springer, 2021, pp. 625–635.

[2] Trainer Hill, “Pokémon tcg tournament metagame aggregates,” https://www.trainerhill.com, 2026, aggregates Limitless TCG tournament results; accessed: 2026-02-20.

[3] Limitless TCG, “Limitless tcg tournament platform,” https://limitlesstcg.com, 2024, accessed: 2026-02-20.

[4] C. E. Shannon, “A mathematical theory of communication,” The Bell System Technical Journal, vol. 27, no. 3, pp. 379–423, 1948.

[5] T. J. Schaefer, “On the complexity of some two-person perfect-information games,” in Proc. 8th ACM Symp. Theory of Computing (STOC), 1978, pp. 216–225.

[6] A. S. Fraenkel and D. Lichtenstein, “Computing a perfect strategy for $n \times n$ chess requires time exponential in n,” Journal of Combinatorial Theory, Series A, vol. 31, no. 2, pp. 199–214, 1981.

[7] M. Bowling, N. Burch, M. Johanson, and O. Tammelin, “Heads-up limit hold’em poker is solved,” Science, vol. 347, no. 6218, pp. 145–149, 2015.

[8] N. Brown and T. Sandholm, “Superhuman AI for heads-up no-limit poker: Libratus beats top professionals,” Science, vol. 359, no. 6374, pp. 418–424, 2018.

[9] ——, “Superhuman AI for multiplayer poker,” Science, vol. 365, no. 6456, pp. 885–890, 2019.

[10] D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou et al., “A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play,” Science, vol. 362, no. 6419, pp. 1140–1144, 2018.

[11] O. Vinyals, I. Babuschkin, W. M. Czarnecki et al., “Grandmaster level in StarCraft II using multi-agent reinforcement learning,” Nature, vol. 575, no. 7782, pp. 350–354, 2019.

[12] P. I. Cowling, C. D. Ward, and E. J. Powley, “Ensemble determinization in Monte Carlo tree search for the imperfect information card game Magic: The gathering,” IEEE Transactions on Computational Intelligence and AI in Games, vol. 4, no. 4, pp. 241–257, 2012.

[13] C. D. Ward and P. I. Cowling, “Monte Carlo search applied to card selection in Magic: The Gathering,” in 2009 IEEE Symposium on Computational Intelligence and Games, 2009, pp. 9–16.

[14] A. Santos, P. Barros, and M. Aragão, “Monte Carlo tree search experiments in Hearthstone,” in Proc. IEEE Conf. Computational Intelligence and Games (CIG), 2017, pp. 272–279.

[15] S. Zhang and M. Buro, “Improving Hearthstone AI by combining MCTS and supervised learning algorithms,” in Proc. AAAI Conf. Artificial Intelligence and Interactive Digital Entertainment (AIIDE), 2017, pp. 68–74.

[16] M. Bjørke and D. Fludal, “Deckbuilding in Magic: The Gathering using a genetic algorithm,” in Proceedings of NISK, 2017.

[17] T. Dockhorn et al., “Introducing the hearthstone-AI competition,” arXiv preprint arXiv:1906.04238, 2019.

[18] J. Kowalski and R. Miernik, “Summarizing strategy card game AI competition,” in 2023 IEEE Conference on Games (CoG), 2023, pp. 1–8.

[19] G. Gonthier, “Formal proof — the four-color theorem,” Notices of the American Mathematical Society, vol. 55, no. 11, pp. 1382–1393, 2008.

[20] J. Avigad, K. Donnelly, D. Gray, and P. Raff, “A formally verified proof of the prime number theorem,” ACM Transactions on Computational Logic, vol. 9, no. 1, pp. 2:1–2:23, 2007.

[21] The mathlib Community, “The Lean mathematical library,” in Proceedings of the 9th ACM SIGPLAN International Conference on Certified Programs and Proofs (CPP). ACM, 2020, pp. 367–381.

[22] T. Hales, M. Adams, G. Bauer, T. D. Dang, J. Harrison, H. Le Truong, C. Kaliszyk, V. Magron, S. McLaughlin, T. T. Nguyen, Q. T. Nguyen, T. Nipkow, S. Obua, J. Pleso, J. Rute, A. Solovyev, A. H. Ta, T. D. Tran, T. H. Trieu, J. Urban, K. Vu, and R. Zumkeller, “A formal proof of the kepler conjecture,” Forum of Mathematics, Pi, vol. 5, p. e2, 2017.

[23] F. Hosch and L. Kovács, “Formalizing hearthstone card effects in Isabelle/HOL,” in Proc. 13th Int. Conf. Interactive Theorem Proving (ITP), ser. LIPIcs, vol. 237, 2022, pp. 1–18, (representative of formal TCG work).

[24] J. M. Smith and G. R. Price, “The logic of animal conflict,” Nature, vol. 246, pp. 15–18, 1973.

[25] P. D. Taylor and L. B. Jonker, “Evolutionary stable strategies and game dynamics,” Mathematical Biosciences, vol. 40, no. 1–2, pp. 145–156, 1978.

[26] J. W. Weibull, Evolutionary Game Theory. MIT Press, 1997.

[27] The Pokémon Company International, “Pokémon trading card game — rules & resources,” https://www.pokemon.com/us/pokemon-tcg/rules, 2024, accessed: 2026-02-20.

[28] ——, “Play pokémon tournament rules handbook,” https://www.pokemon.com/us/play-pokemon/about/tournaments-rules-and-resources, 2024, accessed: 2026-02-20.

[29] A. Tversky and D. Kahneman, “Judgment under uncertainty: Heuristics and biases,” Science, vol. 185, no. 4157, pp. 1124–1131, 1974.

[30] D. Kahneman and A. Tversky, “Prospect theory: An analysis of decision under risk,” Econometrica, vol. 47, no. 2, pp. 263–291, 1979.

[31] A. V. Banerjee, “A simple model of herd behavior,” The Quarterly Journal of Economics, vol. 107, no. 3, pp. 797–817, 1992.

[32] S. Bikhchandani, D. Hirshleifer, and I. Welch, “A theory of fads, fashion, custom, and cultural change as informational cascades,” Journal of Political Economy, vol. 100, no. 5, pp. 992–1026, 1992.

[33] N. Nisan, T. Roughgarden, E. Tardos, and V. V. Vazirani, Eds., Algorithmic Game Theory. Cambridge University Press, 2007.

[34] J. von Neumann, “Zur Theorie der Gesellschaftsspiele,” Mathematische Annalen, vol. 100, no. 1, pp. 295–320, 1928.

[35] J. F. Nash, “Equilibrium points in n-person games,” Proceedings of the National Academy of Sciences, vol. 36, no. 1, pp. 48–49, 1950.

[36] J. Hofbauer and K. Sigmund, Evolutionary Games and Population Dynamics. Cambridge University Press, 1998.

[37] W. H. Sandholm, Population Games and Evolutionary Dynamics. MIT Press, 2010.

[38] O. Romero and I. Millet, “Analysis of swiss-system tournament pairings,” Journal of Quantitative Analysis in Sports, vol. 18, no. 3, pp. 213–229, 2022.

[39] K. Buzzard, J. Commelin, and P. Massot, “Formalising perfectoid spaces,” in Proceedings of the 9th ACM SIGPLAN International Conference on Certified Programs and Proofs (CPP). ACM, 2020, pp. 299–312.