Combining Ink, Water, and Cardboard - Using AI Solutions and Curated Data Sets to Make Clear Solutions For Advanced Data Problem Sets in Magic: The Gathering and Beyond

Andrew Hollis
Emerging Technology Support, LLC
Mooresville, NC
andrew.hollis@emtecsu.com

Christopher Samulski
Argano
Plano, TX
chris.samulski@argano.com

Samantha Sessions
Argano
Plano, TX
sam.sessions@argano.com

Luke Fallis
Argano
Plano, TX
luke.fallis@argano.com

ABSTRACT

Artificial intelligence (AI) is often heralded as a panacea for complex challenges, yet this perception can lead to unclear outcomes. Our more nuanced approach views AI as a cauldron, where, with the right blend of tools and meticulously curated data, it brews a clear, liquid-like solution masterfully crafted by the artisan for perfect water-like clarity instead of one that is murky or cloudy. Our study explores this dynamic through the lens of *Magic: The Gathering's* Commander format, using Wizards of the Coast's newly proposed Bracket system as a baseline. We compare this expert-driven framework against AI-assisted and fully agentic alternatives, evaluating each through Microsoft Azure tools, including Azure Foundry AI, CosmosDB, and Copilot. Crucially, this paper does not argue for AI as a replacement for domain expertise, but rather as a force multiplier—amplifying the precision, fairness, and scalability of human-driven methods when used responsibly. By highlighting the role of human oversight in defining system boundaries, curating inputs, and validating outcomes, we illustrate how collaborative AI workflows align with Microsoft's Responsible AI pillars. A player survey conducted with a newly released preconstructed deck serves to validate which approaches resonate most with real users. Our findings reinforce that the most successful agentic systems are those shaped—and limited—by the steady hand of human judgment.

ABOUT THE AUTHORS

Andrew Hollis is the Special Projects Lead at Emerging Technology Support, LLC, specializing in immersive training simulations, VR/AR systems, agentic AI workflows, curriculum development, on-the-podium teaching, serious games, and as a roleplayer. With over 14 years of experience across defense and enterprise sectors, he has developed cross-platform tools using various game engines and a bevy of programming languages for various computer-based training solutions, as well as generative and agentic AI solutions. He also teaches animation and game development at Regent University. Andrew holds an MFA in Game Design and a BFA in Animation. In addition, as a professional tabletop role-playing game master and player of Magic: The Gathering for almost 30 years, he brings domain expertise in both the game and responsible human-AI collaboration. He has served as a member of the Industry Track committee for MODSIM World since 2024.

Christopher Samulski is Vice President – Delivery, Data & AI at Argano, a digital consultancy delivering high-performance operations through strategic alliances with Microsoft, Salesforce, SAP, and Oracle. He leads solutioning and consulting delivery for Argano's Microsoft Azure Services area, where his team delivers data engineering, analytics, AI applications, agentic automation, and cloud infrastructure services. Christopher began his career in the U.S. Air Force, where he worked with operational data and maintenance systems supporting fighter aircraft—an experience that laid the foundation for his focus on data integrity, systems thinking, and mission-critical operations. He later transitioned to the private sector, advancing through roles as a data analyst, database administrator, and enterprise architect. With over 15 years of experience, he has helped organizations across manufacturing, logistics, media, distribution, life sciences, professional services, and energy industries drive digital and AI transformation. He holds a Bachelor of Science in Management Information Systems from the Rochester Institute of Technology. Outside of work, he is an avid Magic: The Gathering player and tabletop RPG game master.

Samantha Sessions is a Senior Data Analytics Consultant at Argano, where she has worked for the past five years focusing on data management, modeling, and developing data visualizations using Microsoft Fabric. Her work supports Argano's clients in making business decisions based on data analysis. She has previous experience in the consumer packaged goods sector, finance, trade agreements, deductions processing and analysis, and product price list management and rationalization. Samantha holds a Bachelor of Science in Psychology from Old Dominion University. In her free time, she participates in five tabletop role-playing games.

Luke Fallis is an Associate Data Analyst with over two years of experience at Argano. He began his IT career five years ago, working with 3PL and warehousing systems, and later transitioned to a Data Analyst role. He holds a Bachelor's Degree in Animation from Regent University and a Master's Certification in Modeling and Simulation from Old Dominion University. His professional background includes experience in retail and food service operations as well as business logistics across various warehousing platforms. Outside of work, he has participated in Magic: The Gathering and other tabletop games.

Combining Ink, Water, and Cardboard - Using AI Solutions and Curated Data Sets to Make Clear Solutions For Advanced Data Problem Sets in Magic: The Gathering and Beyond

Andrew Hollis
Emerging Technology Support, LLC
Mooresville, NC
andrew.hollis@emtecsu.com

Christopher Samulski
Argano
Plano, TX
chris.samulski@argano.com

Samantha Sessions
Argano
Plano, TX
sam.sessions@argano.com

Luke Fallis
Argano
Plano, TX
luke.fallis@argano.com

INTRODUCTION

Agentic AI has gained attention for its ability to transform static large language models (LLMs) into dynamic task-driven systems. By leveraging tool access, API calls, and procedural logic, these agents can perform increasingly complex tasks, raising hopes about AI's role as a decision-maker in data-heavy environments (IBM, 2024). However, this paper argues that the true power of Agentic AI lies in its role as an augmentative tool rather than a standalone solver. Like a cauldron filled with random ingredients, an AI without guidance produces something messy and unpalatable. Yet with precise ingredients and a master chef's direction, it can become clear and reliable.

AI is often misunderstood as an autonomous solution rather than a tool. This misconception—fueled by both public fear and commercial optimism—has skewed how individuals approach the technology. As Virginia Tech notes, this framing leads some to view AI as a silver bullet for all challenges, while others avoid it entirely due to exaggerated ethical concerns. In reality, AI is akin to a hammer: it requires the right user, the right use case, and the right preparation (Virginia Tech, 2023).

Magic: The Gathering (MTG) is a collectible trading card game first published in 1993 that blends strategic deckbuilding with competitive gameplay across multiple formats. In each game, players use decks composed of spells, creatures, artifacts, and lands to reduce their opponents' life total to zero or fulfill alternate win conditions. One of the most popular variants is the Commander format. In this format, players construct a 100-card singleton deck built around a legendary creature (the "Commander") and limited to the color identity of that Commander. Commander is typically played in multiplayer pods of four, emphasizing long-term strategy, political interaction, and dramatic gameplay (Wizards of the Coast, n.d.).

The format introduces unique rules such as the Commander damage mechanic, command zone access, and the use of singleton card construction (with exceptions for basic lands). Unlike tournament-standard formats, Commander originated as a casual variant emphasizing storytelling, theme, and social contract. Its rules were developed by the Commander Rules Committee, a group of players external to Wizards of the Coast, until the format's governance was partially brought in-house. Today, the Commander Philosophy embraces both competitive and casual play styles while encouraging open discussion about deck power levels—commonly referred to as "Rule 0" conversations (MTGCommander.net, 2024). Originally governed by the independent Commander Rules Committee, the format emphasized informal discussions and pre-game expectations (commonly referred to as Rule 0). Wizards of the Coast now manages the format directly but has maintained this foundational philosophy while seeking to formalize how decks are categorized through tools like the Commander Bracket (MTGCommander.net, n.d.; Wizards of the Coast, 2024).

The complexity of the game itself further amplifies the challenge for AI evaluation. MTG as a whole is Turing complete, meaning it can simulate any computation given enough resources (Churchill et al., 2019). This theoretical property, while not typical of most in-game interactions, underscores the incredible intricacy possible within the rules system. It is not just Commander that is complex—the entire game engine allows for computational universality. This places unique demands on any AI system designed to evaluate gameplay, requiring more than surface-level parsing of cards or decklists.

This paper uses MTG not to solve a gaming problem, but to model a larger insight: effective AI systems must be constrained by human-defined frameworks that promote fairness, maintain reliability, ensure transparency, and assign accountability. Our work in Commander deck classification serves as a microcosm of this broader principle, showing how responsibly designed AI tools, when paired with expert input, can clarify complex data landscapes rather than obscure them.

PROBLEM STATEMENT

Agentic AI systems are only as reliable as the information and instructions provided. When humans are removed from the loop, these systems can begin to drift, producing outputs that are technically correct from a linguistic or logical perspective but misaligned with the context, domain expectations, or user goals. Since the results of an agent are ultimately intended for human consumption, a human must be involved not only in tasking the agent but in designing its workflow, curating its data sources and memory stores, guiding prompt input, and validating its reasoning conclusions.

When working with incomplete or uncurated data, LLMs often produce hallucinations—confident answers that sound plausible but are factually incorrect. This stems from the fact that LLMs are designed to respond to prompts with what they predict is the most acceptable or expected answer, not necessarily the most accurate one. Without clearly defined heuristics and validation protocols, even minor inconsistencies in data can propagate into major errors.

Human-curated methods provide a safeguard against these issues. While they may be less efficient in terms of processing speed, they often yield higher fidelity outputs. Human curators are capable of prioritizing outcomes based on nuance, context, and experience—qualities that are difficult to encode algorithmically. For tasks that involve social interaction, interpretation, or consensus building—as is common in the evaluation of Commander decks—accuracy and clarity outweigh speed and volume.

In Figure 1, the average Power, a game feature of creature cards, which indicates the amount of damage it deals, is compared to the cards' color identity. There seems to be a correlation between the number of colors in a card's identity and the tendency for the card to have a higher power than average. While this is true according to the data, this is not necessarily an indicator of a card being better than another card. However, without knowledge of what indicates what makes a card better than another card in any context, the LLM may choose this metric as an important indicator when determining how a card ranks overall.

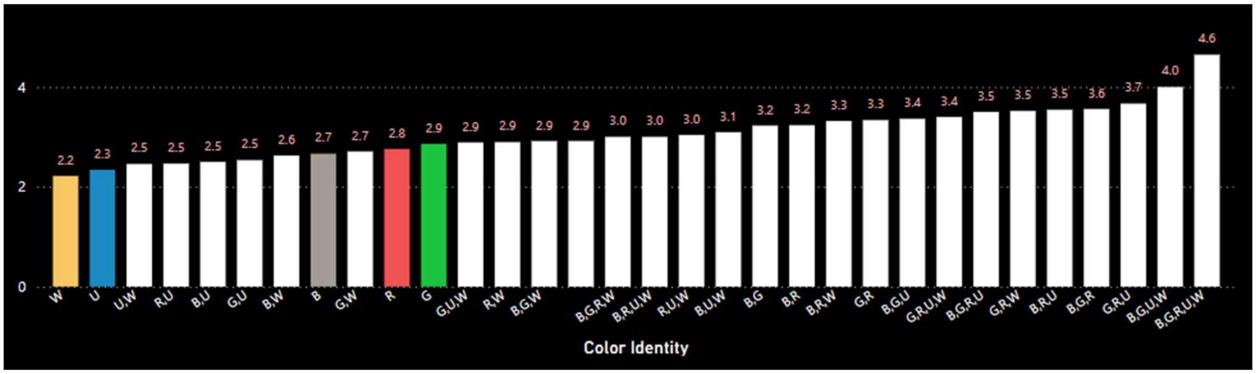


Figure 1. Average Power of Creature Cards by Color Identity

Another indicator featured on every card is the card's rarity. Generally, it would be assumed that the rarer the card, the more powerful the card. This is one of the designed reasons that cards appear higher in rarity. However, the rarity of the top 100 cards in commander decks according to EDHREC, a website that provides data based on decklists published to various websites, is nearly equal between common and rare cards, and uncommon and mythic cards only represent 14% of the remaining cards. Of course, this doesn't mean that the LLM will disregard the reasons cards appear at higher rarities, and will use rarity as a factor in determining a card's overall value to a deck.

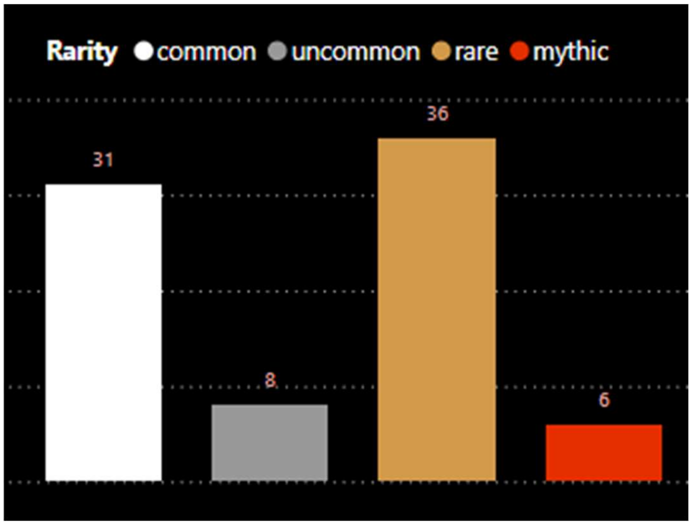


Figure 2. Rarity of the Top 100 Cards in decks according to EDHREC.com

The central hypothesis of this paper is that Agentic AI systems, when guided by human-developed structures, curated datasets, and oversight mechanisms, outperform unguided systems in terms of quality and reliability. Deck rating systems built solely on unstructured inputs or purely autonomous logic lack the embedded game knowledge and social understanding necessary to reflect real-world Commander play.

Microsoft Pillars of Responsible AI

The challenges outlined above—data hallucination, rearing drift, and contextual misalignment—are not specific to Commander deck evaluation. They represent systemic risks in AI deployment whenever systems operate without sufficient human oversight and curation. Microsoft's Responsible AI framework offers six guiding principles to mitigate these risks. Privacy and Security are not included since the agents developed for the context of this paper do not collect personal data and represent a pre-production proof of concept. It is highly recommended that any system brought to production follow the tenets of all five pillars or a similar responsible AI policy framework. The remaining pillars are outlined below.

- Fairness: Human oversight helps identify and correct biases in data and model behavior across similarly situated cases.
- Reliability and Safety: Human involvement ensures systems operate as intended and can handle edge cases or misuse scenarios.
- Inclusiveness: Designing with human context ensures systems support users of diverse backgrounds, needs, and abilities.
- Transparency: Human-guided systems offer traceable logic, making decisions easier to understand and verify.
- Accountability: Humans remain responsible for system outcomes, supported by documented methods and oversight structures.

METHODS AND IMPLEMENTATION

To evaluate the impact of human involvement on agentic AI workflows, we crafted digital twins of human deck evaluators with each following five distinct methods for rating Commander decks. Each method corresponds to a different blend of human input, AI assistance, and agent autonomy. These methods were designed not only to measure deck power but to test the systems themselves, particularly their transparency, reproducibility, and fidelity to the experience of actual play. The five systems are summarized in the following table.

Table 1. Deck Rating Methods - Refer to Appendix A for more Details

Name (Method Number)	Development
WOTC Official Bracket System (1)	Fully Human Crafted by Designers and Experts at WOTC
Human-Evaluated Bracket System (2)	LLM Assisted Improvement to Method 1
Power/Toughness Evaluation System (3)	Semi-Human Crafted by Andrew Hollis for Agentic Integration
Power Level Evaluation System (4)	LLM Assisted Improvement to Method 3
T.R.A.C.K. System (5)	Fully LLM Generated

Appendix A outlines the five Commander deck rating methodologies varying in design philosophy, human involvement, and AI integration. Method 1, developed by Wizards of the Coast, introduces a bracket-based system grounded in philosophy and card usage patterns. Method 2 enhances this using LLM prompts to implement a scoring quiz that maps decks to bracket tiers based on measurable traits. Method 3, designed by Andrew Hollis, introduces Power and Toughness axes informed by Agentic psychographic evaluators and trait scoring to reflect both explosiveness and resilience. Method 4 builds on this by prompting the LLM to simplify the system into five universal traits averaged into a single score. Method 5, the T.R.A.C.K. system, is a fully agent-generated solution that scores decks on Threat, Resource Engine, Anchors, Control, and Kickoff speed, averaging them into an accessible final score.

COMPARISON OF DATA INPUTS AND EVALUATION CRITERIA

The dataset used across all five evaluation methods was the default bulk card data provided by Scryfall, retrieved from the Scryfall website's API documentation as of June 11, 2025, at 09:09 UTC (Scryfall, n.d.). This dataset includes detailed JSON representations of all Magic: The Gathering cards, including metadata relevant to gameplay, card legality, and card mechanics. Before use, the data was filtered to remove cards not legal in the Commander format and to eliminate duplicate printings, retaining only the least expensive legal version of each card. This filtering ensured a standardized and cost-efficient view of the card pool, suitable for computational analysis.

Table 2. Method vs JSON Size

Method(s)	JSON Size
Original Json	~495MB
1 & 2	~7MB
3 & 4	~65MB
5	~156MB

Method 1 and Method 2 utilized a minimal subset of this data, retaining only the card name, oracle text, and a manually curated boolean flag indicating whether the card is a "Game Changer." These methods prioritized interpretability and simplicity, enabling agents to reason over human-defined categories. In contrast, Methods 3 and 4 required a richer data context. Their curated JSON inputs included the card name, oracle text, Game Changer status, mana cost, converted mana cost (CMC), rarity classification, and estimated market price. These expanded inputs enabled nuanced evaluation across gameplay impact, accessibility, and deck construction roles. Method 5, being a purely agent-generated methodology, leveraged the widest data scope. It retained all legal Commander cards from the original

JSON file and required no additional curation beyond ensuring legality and removing duplicates. This resulted in a broader dataset but risked lower interpretability due to the lack of human-imposed constraints.

SYSTEM ARCHITECTURE AND METHODOLOGY

To support these evaluations, a modular system was developed using modern AI infrastructure tools. Visual Studio Code and GitHub Copilot facilitated Python-based development for JSON processing, data filtering, and API integration (GitHub, n.d.; Microsoft, n.d.). CosmosDB served as the primary datastore for card JSON files, decklists, and evaluation results. Azure AI Foundry was used to host and test each agent's performance under structured system prompts. Each evaluation method followed a repeatable architecture:

- Curate a JSON file based on method-specific field requirements.
- Upload curated data to CosmosDB.
- Deploy the associated evaluation agent via Azure AI.
- Submit decklists for analysis and store results.

All agents were issued a task prompt specifying their evaluation methodology, response format, and an example decklist. During runtime, each agent queried the Commander Spellbook API to append combo-related data to cards in real-time to a REST API on the CommanderSpellbook website (Commander Spellbook, n.d.). This ensured analysis reflected not only static traits but dynamic synergies relevant to each card. The orchestrating agent controlled workflow and sequence execution, invoking child agents where applicable and aggregating final scores. This enabled effective handling of token limitations imposed by Azure AI and circumvented restrictions associated with multi-agent nesting in Foundry.

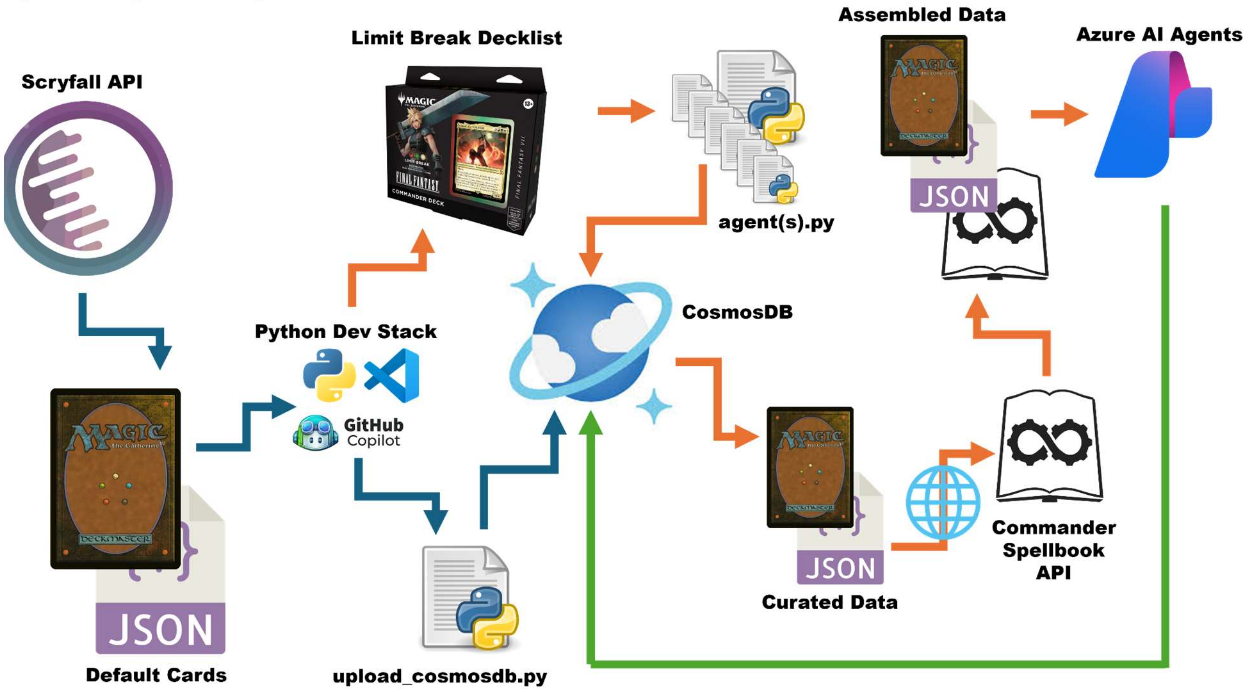


Figure 3. System Architecture Diagram

Agentic Workflow Overview

- Method 1, 2, 4, 5: Used one agent each.
- Method 3: One main agent and five supporting agents (representing player psychographics).
- All results were stored in CosmosDB for retrieval upon duplicate submissions.

Python's modularity and Azure's scalability were critical for executing evaluations iteratively across methods, ensuring consistency in testing while allowing flexible expansion.

COST-BENEFIT ANALYSIS OF HUMAN VS AGENTIC WORKFLOWS

The development of each evaluation method involved varying degrees of human labor, technical expertise, and system design complexity. Method 1 required the most manual effort, as it involved expert-curated heuristics and logic modeling for card evaluation, with an estimated 40–60 hours of work from data architects and game designers. Method 2 reduced this cost by leveraging prompt engineering and minor LLM feedback cycles. Method 3, though AI-driven, required substantial coordination across five psychographic agents and supporting logic, making it the most complex in terms of development. In contrast, Methods 4 and 5 demanded minimal upfront effort—Method 4 required lightweight prompt tuning, and Method 5 was generated almost entirely by an LLM, requiring less than 10 hours of developer input. These distinctions demonstrate that while human-in-the-loop methods are costlier to build, they often yield higher interpretability and control.

Table 3. Estimated Dev Hours per Method

Method(s)	Estimated Dev Time	Primary Roles	Complexity
1	40-60 hrs	Data Architect, Game Designer	High (manual heuristics)
2	20-30 hrs	Prompt Engineer, Data Analyst	Moderate (AI-assisted)
3	60-80 hrs	AI/ML Engineer, Game Designer	Very High (multi-agent)
4	10-15 hrs	Prompt Engineer	Low (simple AI scoring)
5	8-10 hrs	AI/ML Engineer	Very Low (fully LLM-based)

Infrastructure Costs Per Deck

The infrastructure costs for evaluating a single Commander deck using Azure services are remarkably low. Cognitive Services, which hosts the GPT-4 or GPT-4o models through Azure Foundry, contributes approximately $0.045 per deck evaluation. CosmosDB, operating in serverless mode, supports lightweight JSON storage and retrieval with an average cost of just $0.014 per deck. Azure Machine Learning services contribute around $0.005, primarily for workflow testing and metric tracking, while the App Service and associated telemetry incur less than a penny per evaluation. Taken together, the total infrastructure cost per deck remains between $0.06 and $0.07, even when using modern multi-agent orchestration and cloud-native tools. This cost-efficiency enables the system to operate at scale with minimal financial overhead.

Table 4. Infrastructure Cost Per Deck

Service	Cost Per Deck	Purpose
Cognitive Services (AI)	~$0.045	GPT-4 or GPT-4o agent processing
CosmosDB (Serverless)2	~$0.014	Deck storage, card lookups
Azure ML (Testing/Logs)	~$0.005	Evaluation harness testing
App Service + Logs	~$0.001	Foundry agent orchestration + telemetry
Total Estimated Cost	~$0.06–$0.07	Based on actual usage across 100 decks

Scalability Comparison

One of the clearest advantages of agentic workflows is their exceptional scalability. Manual evaluation requires 1 minute per card, resulting in 62 to 100 minutes of labor per deck and costing between $50 and $140, depending on the evaluator's role and hourly rate. This linear cost structure quickly becomes impractical when analyzing large volumes of decks. In contrast, the Azure-hosted agentic methods complete evaluations in 2 to 5 minutes, regardless of deck size, incurring a flat infrastructure cost of less than $0.07 per run. These agent-based methods enable batch processing, parallel execution, and on-demand evaluation, making them ideal for scaling across thousands of decks. The nearly constant evaluation time and cost illustrate the dramatic efficiency gains realized when replacing manual processing with intelligent automation.

Table 5. Human vs Agentic Time/Costs

Evaluator	Time per Deck	Cost Per Deck
Human (Manual)	62-100 min	~$50-$140
Agentic (Azure AI)	2-5 min	~$0.06–$0.07

EXPERIMENT AND RESULTS

To conduct the experiment that evaluated the results of the deck rating Agents, a deck was chosen that's contained enough new cards to ensure that our data set provided was the only source of most of the deck's information instead of what could be already part of the LLM used by our agents which was GPT-4o, which at the time of use did not have access to the information about our chosen deck. The chosen deck was the pre-constructed deck created for sale by Wizards of the Coast called "Limit Break," which was released with the Final Fantasy set on June 13th, 2025. As an important note before the results, Wizards intends their pre-constructed decks to fall under Bracket 2 of their deck rating methodology, as this is what they considered the reference point for what a Bracket 2 deck experience should be.

Table 6. Deck Rating Results - Appendix B contains the Summarized Agent Output from the Experiment

Method	Rating	Accuracy Result
Method 1	Bracket 2	Accurate (Intended)
Method 2	Score 6 - Bracket 1	Underestimation
Method 3	Power/Toughness 7/6	Overestimation
Method 4	Power Level 7	Overestimation
Method 5	T.R.A.C.K - 6.5	Overestimation

The evaluation results demonstrate that only Method 1 produced the intended outcome, correctly placing the deck in Bracket 2. Method 2 significantly underestimated the deck, while Methods 3, 4, and 5 consistently overestimated its power. This highlights that, although each agent was capable of generating a complete rating, only the system developed by Wizards of the Coast aligned with the expected classification. The discrepancies observed in the other methods do not imply their evaluations were fundamentally flawed; rather, they suggest that the underlying data, methodologies, or system prompts were insufficiently calibrated. These methods require further refinement to produce accurate and reliable results consistently.

Survey on the Results and Insights

To evaluate the effectiveness and preference of the five proposed Commander deck evaluation methods, a survey was conducted with a targeted group of experienced MTG Commander players. A total of seven responses were received, each including at least one preferred and one least preferred method, along with optional demographic information. The survey aimed to understand how well the results of each agent and the methods used aligned with community expectations and whether the methodology was comprehensible, intuitive, and aligned with perceived deck strength. All seven respondents identified themselves as regular Commander players, with most playing weekly. The ages of the participants ranged from 18 to 54 years, encompassing a broad generational spectrum. Player psychographics varied widely, with most respondents selecting multiple roles such as Spike, Timmy, Mel, and Johnny, indicating well-rounded experience and varied priorities. The survey responses revealed several key insights. Method 1 emerged as the most preferred approach, chosen by four out of seven participants. It was commended for its alignment with the existing bracket framework and for producing evaluations that matched expectations for preconstructed decks. Method 2 proved to be the most polarizing, with some respondents appreciating its nuanced assessment of a deck's mediocrity, while others criticized it for significantly underestimating the deck's actual strength. Method 4 received the most negative feedback, with three participants ranking it as their least favorite due to its confusing, inaccurate, or overly generous numerical outputs.

Meanwhile, Methods 3 and 5 garnered limited but positive attention, particularly from players who valued creativity or flavor-driven deck design. Player comments reflected a general mistrust of AI-generated numerical ratings without transparent justification. Human-curated systems were seen as more credible, particularly when their logic mapped clearly onto the known Commander Bracket system. However, niche preferences also supported methods that recognized non-meta aspects like flavor, creativity, and combat-centric playstyles. The survey results met our expectations for a preference in an Agentic result, whose methodology was crafted by a human expert.

The Methods versus the Pillars

The methods were not at any point given instructions on ensuring they follow the pillars set forth by Microsoft for responsible AI. Method 1 fully meets almost all of Microsoft's Responsible AI pillars through structured guidance, expert curation, inclusive design, and institutional accountability, with only some issues with transparency for top tiers. Method 2 aligns closely in fairness and transparency with its point-based rubric but lacks oversight, making it vulnerable to player bias. Method 3 introduces psychographic agents to capture diverse player styles, enhancing inclusiveness but reducing clarity and reliability due to complexity and opaque logic. Method 4 simplifies the Power/Toughness concept into an accessible 1–10 scale, improving usability and consistency, though it sacrifices accountability and formal validation. Method 5, the T.R.A.C.K. system, is entirely autonomous and offers accessible scoring but fails to meet critical pillars due to the absence of human oversight, bias mitigation, or safeguard mechanisms.

Table 7. Comparison Table: Alignment with Responsible AI Pillars

Method	Fairness	Reliability & Safety	Inclusiveness	Transparency	Accountability
Method 1	✓	✓	✓	-	✓
Method 2	✓	X	✓	✓	-
Method 3	✓	X	✓	-	X
Method 4	X	X	-	-	-
Method 5	X	X	X	X	X

DISCUSSION

The findings of this study yielded several significant insights into the capabilities and limitations of both human-guided and agentic AI methodologies in Commander deck evaluation. As expected, Method 1—grounded in human expertise and supported by curated data—delivered the desired outcome. Its success underscores the value of explicit domain knowledge and structured logic in augmenting AI performance. Method 2, which expanded on Method 1 via LLM guidance, underestimated the power level of the test deck. Methods 3 through 5, despite diverse approaches and deeper trait modeling, tended to overestimate the deck's strength.

Method 1 outperformed others due to the alignment between its underlying human-devised framework and its precisely filtered data. The ability to provide clear expectations and verification paths enabled a high degree of trust and reproducibility. Methods 3, 4, and 5, all of which relied heavily or exclusively on agentic AI, exhibited several constraints. Chief among them was the lack of comparative context—agents were tasked with rating cards without the means to benchmark traits against a structured reference set. This led to inflated or inconsistent ratings when cards were analyzed in isolation. Moreover, some methods lacked access to normalized metadata such as tier lists or pre-classified game changers, limiting their calibration capabilities (Wizards of the Coast, 2025).

Human involvement proved critical in reducing computational complexity and memory overhead. Expert-curated methodologies produced leaner datasets, and evaluations were more interpretable and computationally efficient. Most notably, human oversight enabled systems to provide outputs that users could verify and trust—an essential feature in communal formats like Commander. The results highlight a broader implication: current agentic AI systems, while promising, are most effective when used in tandem with expert-designed data and logic scaffolding. Rather than replacing domain experts, agentic AI is better viewed as an augmentation layer—powerful in handling volume and nuance, but bounded by the quality of its instructional framework. This study affirms the enduring relevance of human experts in AI-driven workflows. As agents take on greater roles in analysis, experts must continue to design, oversee, and validate these systems. Simplistic automation may appear cost-effective, but true reliability and domain alignment emerge only through expert stewardship. Consequently, future AI deployments should prioritize symbiotic human-agent design, with clearly defined boundaries of agency.

BEYOND DECK RATING

The findings of this study reveal more than a classification challenge in tabletop gaming—they expose a structural need across industries for human-integrated agentic AI workflows. While agentic systems offer unprecedented speed, scale, and adaptability, their full potential is only realized when paired with expert-driven constraints, curated datasets, and contextual oversight. This fusion is not just advantageous; it is essential for trust, performance, and responsible deployment. Across all applications, this research reinforces one central idea: Agentic AI must remain grounded in human understanding, intention, and accountability. Rather than displacing domain expertise, agentic tools should act as accelerators for it.

Digital Wargame Balance and Decision Modeling

In digital wargames—especially those with training or defense applications—outcomes must reflect not just logic, but intent. Fairness and balance are human judgments that emerge from goals, doctrines, and historical precedent. Agentic tools can simulate thousands of scenarios, but without human calibration and interpretation, their outcomes risk reinforcing or creating biases. Incorporating human control loops ensures Fairness and Reliability, allowing these models to reflect real-world asymmetry without losing strategic fidelity. In this context, human integration becomes a requirement—not a luxury—for mission-ready simulation environments.

Agent Behavior in Military Training Sims

Modeling agent behaviors in simulations is as much a behavioral science as a technical challenge. Realistic opponent or ally behavior must reflect cultural, tactical, and environmental factors—dimensions that resist codification. Agentic systems can parameterize behavior, but without humans setting the frame, results drift. Integrating human experts ensures Accountability for decisions made in virtual environments and supports Transparency, where trainees and developers can understand not just what agents did, but why. As simulations become more immersive, this clarity becomes a cornerstone of training credibility and effectiveness.

Human-AI Classification in High-Stakes Domains

Agentic classification systems are increasingly being used in finance, medicine, and security. But trust in classification demands more than high throughput—it requires explainability and inclusive design. Just as MTG players demand to know why their deck was rated a certain way, patients and auditors must understand how an AI reached a diagnosis or flagged a transaction. Human integration ensures Transparency and Inclusiveness, guarding against opaque or discriminatory logic. Classification without context is not insight—it’s risk.

Intelligent Customer Support Prioritization and Resolution Routing (Argano Customer Solution)

In high-volume customer support environments, agentic AI systems illustrate the need for human-in-the-loop workflows across critical service functions. A global telecommunications provider faced delays and inconsistency in triaging customer issues across multiple channels. Traditional automation failed to distinguish between nuanced scenarios, such as an emotionally distressed VIP with a minor issue versus a calm report of a major outage.

To address this, a collaborative agent architecture was deployed: one agent assessed tone and urgency, another ranked customer value and support history, and a third orchestrated routing decisions. Low-priority tickets were handled by a generative agent, while high-severity cases were escalated to human agents.

This real-time, multi-agent approach exemplifies how agentic AI can scale support efficiently without abandoning contextual nuance. But critically, the system's effectiveness hinged on human-curated logic and oversight. In line with findings from deck classification and simulation modeling, this use case reinforces that agentic tools must be guided by domain expertise to ensure Fairness, Transparency, and Accountability, especially in emotionally sensitive or high-stakes contexts. AI here doesn't replace the human—it enables better, faster human-informed action.

Monetization and Hobbyist Ecosystems

Even in commercial or recreational ecosystems, human-centered AI enables smarter, fairer engagement. A deck recommendation engine powered by agentic classification becomes meaningful only when it respects budget constraints, player intentions, and community norms—none of which are captured in raw card data. Human-guided agentic systems allow retailers and platforms to recommend upgrades, provide content alignment, and predict trends without alienating new or underrepresented players. Here, Fairness, Inclusiveness, and Accountability ensure that automation supports community growth, not just consumption.

CONCLUSION

Agentic AI is limited by its inability to acquire information in a manner that would be consistent with a human expert's abilities. More data does not always equate to a better result, and a more complicated system that should lead to a more explicit result increases the burden for the human expert to facilitate and task the Agent. Our study showed that the fully human-designed system performed as expected with the data provided. When agentic systems operate under clearly defined human scaffolding, they don't just perform—they earn trust. And in environments as complex as wargames, training simulations, finance, or even a 100-card deck, trust is the most important output of all. The system crafted by the experts also met the pillars of responsibility set forth by Microsoft because it focused on creating a method that a human could perform independently of an agent. We congratulate the team that designed the current Magic Bracket system on this (Wizards of the Coast, 2025). Future developers, analysts, and researchers into Agentic AI should consider whether an Agent has all the information it needs, and only the information it needs, as well as the proper instructions for how to implement that data in its methods. We believe that any of the systems provided could be used as a deck rating system, but the work would need to be done to curate the data, and this delves into something that can be left up to a hobbyist or WOTC to do themselves. We would, however, be happy to share with WOTC our more detailed files and methods should they be willing to see them for help with maintaining their current system.

REFERENCES

Churchill, A., Biderman, S., & Herrick, A. (2019). Magic: The Gathering is Turing Complete. *ArXiv*. https://arxiv.org/abs/1904.09828

Commander Rules Committee. (n.d.). *The philosophy of Commander*. MTGCommander.net. https://mtgcommander.net/index.php/the-philosophy-of-commander/

Commander Spellbook. (n.d.). *Commander Spellbook API*. https://backend.commanderspellbook.com/

GitHub. (n.d.). *GitHub Copilot documentation*. https://docs.github.com/en/copilot

IBM. (n.d.). *Agentic AI*. IBM Think. https://www.ibm.com/think/insights/agentic-ai

Microsoft. (n.d.). *Azure AI Foundry documentation*. Microsoft Learn. https://learn.microsoft.com/en-us/azure/ai-foundry/

Microsoft. (n.d.). *Azure Cosmos DB documentation*. Microsoft Learn. https://learn.microsoft.com/en-us/azure/cosmos-db/

Microsoft. (n.d.). *Python in Visual Studio Code*. Visual Studio Code Docs. https://code.visualstudio.com/docs/languages/python

Scryfall. (n.d.). *Bulk data API*. https://scryfall.com/docs/api/bulk-data

Virginia Tech College of Engineering. (2023, Fall). *Artificial intelligence and the future of engineering*. https://eng.vt.edu/magazine/stories/fall-2023/ai.html

Wizards of the Coast. (n.d.). *Commander format*. Magic: The Gathering. https://magic.wizards.com/en/formats/commander

Wizards of the Coast. (n.d.). *Introduction to Magic: The Gathering*. https://magic.wizards.com/en/intro

Wizards of the Coast. (2024, October 22). *Introducing the Commander Format Panel*. https://magic.wizards.com/en/news/announcements/introducing-the-commander-format-panel

Wizards of the Coast. (2024, April 22). *Introducing Commander Brackets Beta*. https://magic.wizards.com/en/news/announcements/introducing-commander-brackets-beta

Wizards of the Coast. (2025, April 22). *Commander Brackets Beta update*. https://magic.wizards.com/en/news/announcements/commander-brackets-beta-update-april-22-2025

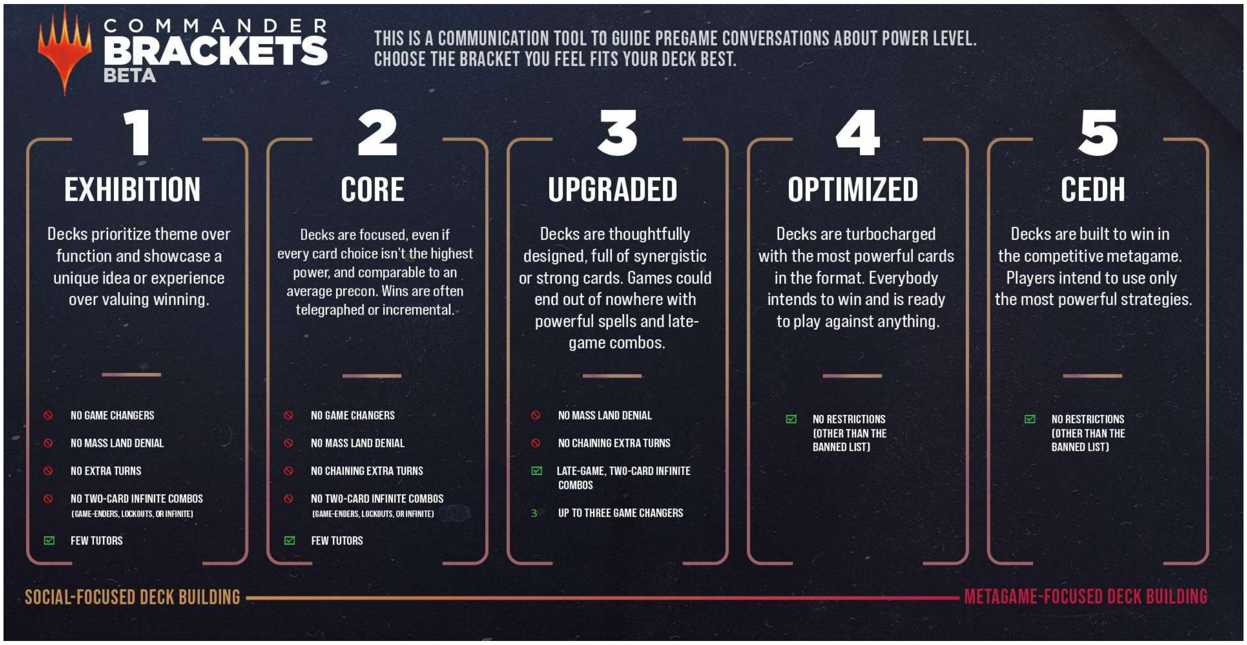
Wizards of the Coast. (2024, September 30). *On the future of Commander*. https://magic.wizards.com/en/news/announcements/on-the-future-of-commander

APPENDIX A - BREAKDOWN OF DECK RATING METHODOLOGIES

Method 1: Wizards of the Coast (WotC) Official Bracket System
Wizards of the Coast (WotC) introduced the Commander Bracket as a formal system to support "Rule 0" conversations by offering structured guidance for categorizing deck power levels. This system divides decks into five brackets:

- Bracket 1 – Exhibition: Designed for decks built around themes or novelty, not focused on victory. They typically exclude infinite combos, mass land destruction (MLD), chained extra turns, and tutors.
- Bracket 2 – Core: Intended to replicate preconstructed (precon) deck experiences. Swingy but still limited in terms of combo density, recursion, and tutoring.
- Bracket 3 – Upgraded: Tuned decks with moderate optimization and the inclusion of a few game-changing cards. Infinite combos are discouraged, but some high-impact cards may be present.
- Bracket 4 – Optimized: Highly tuned decks with extensive synergy and win conditions. Infinite combos, chained extra turns, and heavy tutoring are all permitted.
- Bracket 5 – Competitive EDH (cEDH): Tournament-level optimization with no room for fluff. Every card contributes directly to the win strategy, and efficiency is maximized across the board.

Wizards maintains a curated list of "game changer" cards. This list evolves to accommodate new releases and shifting metas, ensuring consistency across deck evaluations.



COMMANDER BRACKETS BETA

THIS IS A COMMUNICATION TOOL TO GUIDE PREGAME CONVERSATIONS ABOUT POWER LEVEL.
CHOOSE THE BRACKET YOU FEEL FITS YOUR DECK BEST.

1 EXHIBITION

Decks prioritize theme over function and showcase a unique idea or experience over valuing winning.

- ⊘ NO GAME CHANGERS
- ⊘ NO MASS LAND DENIAL
- ⊘ NO EXTRA TURNS
- ⊘ NO TWO-CARD INFINITE COMBOS (GAME-ENDERS, LOCKOUTS, OR INFINITE)
- ☑ FEW TUTORS

2 CORE

Decks are focused, even if every card choice isn't the highest power, and comparable to an average precon. Wins are often telegraphed or incremental.

- ⊘ NO GAME CHANGERS
- ⊘ NO MASS LAND DENIAL
- ⊘ NO CHAINING EXTRA TURNS
- ⊘ NO TWO-CARD INFINITE COMBOS (GAME-ENDERS, LOCKOUTS, OR INFINITE)
- ☑ FEW TUTORS

3 UPGRADED

Decks are thoughtfully designed, full of synergistic or strong cards. Games could end out of nowhere with powerful spells and late-game combos.

- ⊘ NO MASS LAND DENIAL
- ⊘ NO CHAINING EXTRA TURNS
- ☑ LATE-GAME, TWO-CARD INFINITE COMBOS
- 3 UP TO THREE GAME CHANGERS

4 OPTIMIZED

Decks are turbocharged with the most powerful cards in the format. Everybody intends to win and is ready to play against anything.

- ☑ NO RESTRICTIONS (OTHER THAN THE BANNED LIST)

5 CEDH

Decks are built to win in the competitive metagame. Players intend to use only the most powerful strategies.

- ☑ NO RESTRICTIONS (OTHER THAN THE BANNED LIST)

SOCIAL-FOCUSED DECK BUILDING — METAGAME-FOCUSED DECK BUILDING

Figure Appendix-A-1. WOTC Commander Brackets BETA (Current as of June 2025)

Method 2: Human-Evaluated Bracket System (Method 1 LLM-Assisted Improvement)
Method 2 was the result of prompting an LLM with Method 1, the rules of Magic: The Gathering, and example data with the following prompt:

"I want to improve the Bracket System. I've uploaded the comprehensive rules of Magic, as well as a json file that contains specific types of details that may be a part of an individual card. Using all this information as well as the justification for the current Bracket System, write a new Bracket System that improves upon the old and a script that will evaluate a decklist."

The results were a method for players to self-identify their deck's bracket based on card traits and construction, and after evaluating their deck using a simple math equation to determine where their deck fell into the five different brackets presented. Each deck is now analyzed for the following:

- Combo Density
- Tutor Count:
- Extra Turns
- Mass Land Denial
- Game Changers
- Speed
- Intent (Synergy)

A simple quiz awards some points for the deck for each trait, and the deck falls into one of the following brackets:

- Bracket 1 – Theme: 0–8 points
- Bracket 2 – Core: 9–14 points
- Bracket 3 – Synergy+: 15–21 points
- Bracket 4 – Optimized: 22–29 points
- Bracket 5 – cEDH: 30+ points

Method 3: Power/Toughness Evaluation System (Human-Curated with Psychographic Agents)
This method, designed by Andrew Hollis, incorporates an algorithmic and Agentic solution to rating a deck. The system uses one primary agent and five additional agents (Vorthos, Mel, Timmy, Johnny, and Spike) to study certain aspects of the cards. There are five Agents named for certain play styles and demographics, which WOTC has labeled for certain members of its player base. To classify Commander decks using a familiar Magic: the Gathering mechanic—Power and Toughness—as a clear, two-axis rating for the deck's explosive potential (Power) and its resilience and interaction (Toughness).

Each card in a Commander deck is scored across 10 traits, divided into:

- Psychographic Quotient - The value of the card on the play style spectrum, assigned via the Psychographic Agents' voting.
- Storm Scale Inversion - A value assigned to a card based on the highest scoring mechanic on the card, based on Wizards of the Coast's storm scale.
- Game Changer Potential - A value assigned based on whether the card is considered a game changer by WOTC.
- Card Accessibility Index - A composite value of price, print count, and reserved list status.
- Synergy Density - A value based on known archetypes or engines.
- Rarity Impact - A value tied to the assigned rarity of the card.
- Performance Traits
  - Power
    - Tempo Impact - Measures early or explosive play potential.
    - Combo Enabling - Does the card appear in combos in the deck
  - Toughness
    - Interaction Potential - Can the card disrupt, counter, or deter opponents?
    - Resilience & Recursion - Can it protect itself or return/reuse key cards?

This method intends to give players more control over determining results without an Agent's help, and is a nod to the old method of rating decks on a scale of 1–10 before WOTC's system. Once all the cards have been assessed, they are then sorted into two piles based on how high they score in the two performance traits, and then the summation of each pile is divided by the maximum value to give a Power/Toughness rating for the deck.

Method 4: Power Level Evaluation System (Method 3 LLM-Assisted Improvement)
Method 4 was the result of prompting an LLM with Method 3, the rules of Magic: The Gathering, and example data with the following prompt:

"Hello I created a deck rating system that follows an Agentic AI and algorithmic approach, but I want the method to be more accessible to players while still feeling similar to the current approach and not requiring deep mathematical

calculations or evaluations. Update the following methodology with this intent in mind, and edit my code to follow this intent." Followed by Method 3.

Instead of assigning two separate axes (Power and Toughness), this system uses a composite score derived from five universal traits, each scored 1–10 per card. The final deck score is calculated by averaging the total score across the full decklist and normalizing it on a 1–10 scale, similar to the older method players would assess their decks before Method 1's introduction by WOTC. Each card is scored from 1 to 10 in the following five traits:

- Tempo & Speed: How quickly and efficiently a card affects the board or progresses your game plan. Fast mana, haste, ramp, or low-cost value spells score higher.
- Interaction: The card's ability to control, delay, or punish opponents. Includes removal, counterspells, stax, discard, theft, and protective effects.
- Recursion & Resilience: How well a card helps the deck recover, protect key pieces, or maintain presence over time. Includes graveyard recursion, indestructibility, and redundancy.
- Combo & Synergy: The card's ability to enable combos, loops, or deeply synergize with the deck's strategy and commander. Tutors, payoff pieces, and combo glue all rate highly.
- Impact & Uniqueness: How powerful, rare, or central the card is to winning or shifting the game state. Game-changers, alt-win enablers, and "must-answer" threats score highly.

Score Interpretation:

- 1.0–3.0: Casual, inconsistent, or highly experimental deck. Likely lacks interaction or win conditions.
- 3.1–5.5: Thematic or budget-friendly build with limited optimization. Playable but fragile.
- 5.6–7.5: Well-rounded and reliable. Synergistic mid-tier deck with strong combos or interaction.
- 7.6–9.0: Highly optimized and dangerous. Competitive in most pods.
- 9.1–10.0: cEDH or near-cEDH level power. Extremely fast, consistent, and hard to stop.

Method 5: T.R.A.C.K. System (Fully Agentic LLM-Generated Method)
This method was entirely crafted by the LLM using only a decklist, the rules of Magic: The Gathering, and a single JSON file showing the card data it would be provided. T.R.A.C.K. stands for:

- Threat Level - How quickly and dangerously this deck can end the game or dominate the board.
- Resource Engine - How well the deck draws cards, generates mana, and maintains value.
- Anchors & Combos - How many strong synergies, infinite combos, or recursive loops the deck supports.
- Control Tools - How much the deck disrupts or controls opponents (counters, removal, stax).
- Kickoff Speed - How fast the deck starts enacting its strategy.

This system assigns a 1–10 score for each category and averages them to form a final deck score (1.0–10.0). It works for deck creators and opponents to communicate clearly and allows quick comparisons using accessible concepts. Final Power Score = (T + R + A + C + K) / 5 Rounded to the nearest half-point (e.g., 5.5, 7.0, 8.5)

Table Appendix A-1. T.R.A.C.K Power Summary

T.R.A.C.K Power Score	Tier	Summary
1.0-3.0	Casual	Precon-level or janky fun
3.5-5.0	Upgraded	Budget upgrades, casual combos
5.5-6.5	High Casual	Well-tuned but not oppressive
7.0-8.0	Optimized	Near-cEDH power, fast combos
8.5-10.0	Competitive	cEDH territory, 1–3 turn wins

APPENDIX B - DECK EVALUATION FOR THE DECK "LIMIT BREAK"

This Cloud, Ex-SOLDIER deck is a well-rounded, high-casual Voltron/Equipment build with a strong focus on legendary creatures and Equipment synergies. It ramps well, draws plenty of cards, and can quickly become threatening by suiting up Cloud or other legends. While it lacks infinite combos or competitive-level engines, it has enough removal and board wipes to interact with opponents and keep the board clear. The deck's main weakness is the lack of instant-win combos or hard locks, but it makes up for it with consistent value and the ability to pressure opponents through combat. Great for casual to high-casual tables, especially for fans of legendary creatures and Equipment strategies!

Method 1:
Bracket: 2 – Core
This equipment-focused Commander deck fits in the "Core" bracket, meaning it is comparable to a modern precon with some upgrades. It avoids powerful staples such as tutors, infinite combos, extra turns, or mass land destruction, which keeps the power level approachable for most casual tables. The synergy around legendary creatures and equipment is strong, but not oppressive or tuned for competitive play.

Method 2:
Score: 6 (Trait Breakdown: Combo: 0, Tutors: 0, Extra Turns: 0, MLD: 0, Game Changers: 2, Synergy/Speed: 4) — Bracket: 1 – Theme
This deck scores a 6 on the Commander Bracket Classifier rubric, placing it in the "Theme" bracket due to its strong flavor and synergy but absence of high-powered elements. The deck is highly synergistic around equipment and legendary creatures, with a cohesive Final Fantasy VII theme that drives the card choices and strategy. While it includes a few powerful board wipes and synergy engines, the absence of combos, tutors, or cEDH staples keeps it firmly in the casual, thematic range.

Method 3:
Deck Power/Toughness: 7 / 6
This deck is a proactive, midrange build with strong synergy between equipment and legendary creatures, earning it a 7 in power for its ability to create powerful board states and apply pressure. Its toughness is a 6, reflecting decent protection and resilience, but with some vulnerability to repeated wipes and limited recursion. The deck is best for players who enjoy building value and synergy over time rather than relying on combo finishes or extreme resilience.

Method 4:
Power Level = 7
The deck is rated a 7 out of 10 for its balance of strength, synergy, and interactivity. It excels at leveraging legendary and equipment synergies, ramping efficiently, and maintaining board presence, but it does not reach the higher echelons of Commander power due to its lack of infinite combos, stax elements, or oppressive lock pieces. This rating makes it ideal for "high casual" to mid-power tables where value and interactivity are prized over raw efficiency or degeneracy.

Method 5:
T.R.A.C.K = 7.0, 7.5, 5.0, 6.5, 6.5
The T.R.A.C.K. system scores the deck as follows: Threat Level (7), Resource Engine (7.5), Anchors & Combos (5), Control Tools (6.5), and Kickoff Speed (6.5), for an overall Power Score of 6.5. This reflects a well-rounded, high-casual deck with strong resource generation, good board presence, and solid but not overwhelming speed and interaction. The main limitation is in combo potential, but the deck shines in its ability to pressure opponents and recover from disruption, making it a strong choice for interactive yet casual games.

Summary:
All five methods converge on the conclusion that this is a high-casual, synergy-driven Commander deck focused on legendary creatures and equipment, with strong flavor and value-oriented play. It is not competitive or combo-centric but offers plenty of interaction, resilience, and thematic fun for most Commander tables. Use the breakdowns above to determine which evaluation best matches your own play philosophy!

GLOSSARY OF MAGIC TERMS

cEDH (Competitive EDH) – A subformat of Commander where decks are optimized for maximum efficiency, often ending games within a few turns using tight combo lines.

Color Identity – The combination of mana symbols in a Commander's casting cost and rules text, which determines what cards are legal in that deck.

Commander – A legendary creature that serves as the centerpiece of a Commander deck and defines its color identity. Always accessible from the Command Zone.

Command Zone – A game zone separate from the battlefield or graveyard where the Commander resides and can be cast from during the game.

Combo Density – A trait used in deck evaluation methods that measures the number or frequency of infinite or semi-infinite card interactions that result in game-ending sequences.

Converted Mana Cost (CMC) – The total cost to cast a card, formerly a standalone term, now more commonly referred to as "mana value."

Decklist – A complete list of the cards in a specific deck used for evaluation or gameplay.

Game Changer – A designation for cards that can swing or dominate a game, typically flagged manually or by heuristics in deck evaluation methods.

Kickoff Speed – In the T.R.A.C.K. system, this measures how quickly a deck can begin enacting its strategy after the game starts.

Legendary Creature – A creature card with the "Legendary" supertype. Only one of each name can be on the battlefield at a time per player.

Mana – The primary resource in MTG used to cast spells. Comes in five colors and is produced by lands and other sources.

Mass Land Destruction (MLD) – Cards or strategies that destroy multiple lands, often used in high-power or control-oriented decks.

Oracle Text – The most up-to-date and official rules text of a card, used in evaluation systems to determine function and interactions.

Player Archetypes – Common categories used by WOTC to describe player preferences:

- Spike – Values winning through optimization and efficiency.
- Timmy/Tammy – Enjoys big plays, powerful creatures, and dramatic moments.
- Johnny/Jenny – Loves combo creativity and unique deckbuilding.
- Vorthos – Cares deeply about story, lore, and card aesthetics.
- Mel/Melvin – Enjoys mechanical design and elegant interactions.

Power / Toughness (Deck Evaluation) – A metaphorical extension of MTG's creature stats used in Method 3 to represent a deck's offensive capability (Power) and resilience/control (Toughness).

Preconstructed Deck (Precon) – A ready-to-play deck sold by WOTC, typically balanced around Bracket 2 and intended for casual play.

Psychographic Agent – An AI agent representing a player archetype used in Method 3 to simulate how different player types evaluate cards.

Reserved List – A list of cards maintained by Wizards of the Coast that will never be reprinted to protect the long-term value of collectors' investments. Cards on this list often have limited print runs and are difficult to access, influencing the Card Accessibility Index in evaluation systems.

Rule 0 – A core philosophy of Commander that encourages pre-game discussions about power levels, proxies, banned cards, and house rules to ensure a fun and balanced experience for all players.

Scryfall – A comprehensive third-party MTG card database with an open API, used in this paper to source all card data and metadata.

Singleton – A format rule requiring no duplicate non-basic cards. In Commander, only one copy of each non-basic land or card is allowed.

Storm Scale – A scale used by WOTC to assess the likelihood of a mechanic returning in future sets, inverted in Method 3 to evaluate complexity and power.

Toughness – In game terms, the amount of damage a creature can sustain. Method 3 refers to a deck's defensive capability and resilience.

Tutors – Cards that search the library for specific cards, often considered powerful and indicative of high optimization.