

Defining your Problem and Audience

Description of Problem

I am a big fan of Magic: The Gathering. One of the biggest flaws of the game is as the games get more complex, the HUGE amount of rules becomes overwhelming and almost nobody knows how certain interactions work, so certain scenarios take forever to solve rules-wise.

Why This is a Problem

This slows gameplay, creates irritation and arguments, and there are too many rules AND too many cards to easily look up how any given interaction is supposed to play out in a single google search.

Propose a Solution

Propose a Solution

JudgeBot! A bot that uses RAG to retrieve official rules docs and card text/card rules, augments them into necessary information to solve a given situation, and then generates the solution, aka the official ruling!

Describe Tools You Plan to Use

Some kind of Magic the Gathering API, Wizards of the Coast official Rules Documentation, a GPT model, RAGAS for optimization, LangChain/LangGraph.

When Will You Use an Agent? What is the Reason for “Agentic Reasoning” in the app?

Multi-step rules reasoning, Dynamically choose what tools to use next, autonomy, long-term improvement(?), plan sub-tasks.

Dealing with the Data

Describe all of your data sources and external APIs, and describe what you'll use them for.

Scryfall API for oracle text, rulings, card images, legalities, keywords.


<https://magic.wizards.com/en/rules> for a comprehensive document of every rule.

Describe the default chunking strategy that you will use. Why did you make this decision?

This project is going to use a rule-aware chunking strategy specifically designed for Magic: The Gathering Comprehensive Rules. It will first split the text on rule headers (e.g., "405." or "100.6a") to preserve semantic boundaries, and for sections exceeding 800 characters, it will use RecursiveCharacterTextSplitter with a 150-character overlap and hierarchical separators (double newlines, newlines, periods, spaces). Each chunk will be enriched with extracted metadata including rule numbers, titles, and keywords (like "stack", "priority", "combat") to improve retrieval accuracy.

[Optional] Will you need specific data for any other part of your application? If so, explain.

Building a Quick End-to-End Prototype

Build an end-to-end prototype and deploy to local host with a front end (Vercel deployment not required). 

Creating a Golden Test Data Set

Assess your pipeline using the RAGAS framework including key metrics faithfulness, response relevance, context precision, and context recall. Provide a table of your output results.

```
question_id,difficulty,question_short,context_recall,faithfulness,factual_
correctness(mode=f1),answer_relevancy,context_entity_recall,noise_sensitiv
ity(mode=relevant),context_precision,overall_score,quality
1,easy,What does Lightning Bolt
do?,1.0,1.0,1.0,0.937,0.667,0.0,1.0,0.801,High
2,easy,What is the mana cost of
Counterspell?,1.0,1.0,1.0,1.0,0.333,0.0,1.0,0.762,High
3,easy,What type of card is Sol
Ring?,1.0,1.0,0.5,1.0,0.5,0.0,1.0,0.714,High
4,hard,"If my opponent controls Blood Artist, and one of
m...",1.0,1.0,0.86,0.921,0.25,0.0,1.0,0.719,High
5,hard,If my opponent has Doubling Season and I cast a
pl...,1.0,0.667,0.57,0.815,0.5,0.333,1.0,0.698,Medium
6,medium,How do Rest in Peace and Leyline of the Void
inter...,0.75,1.0,0.67,0.913,0.667,0.375,1.0,0.768,High
7,easy,If I cast Counterspell targeting my opponent's
Lig...,1.0,0.0,0.67,0.959,0.5,0.0,1.0,0.59,Medium
8,medium,Can I use Path to Exile on my own
creature?,0.0,1.0,0.86,0.976,0.25,0.25,1.0,0.619,Medium
9,medium,How does the stack
work?,0.667,1.0,0.86,0.708,0.167,0.143,1.0,0.649,Medium
10,medium,What are state-based
actions?,0.5,0.667,0.77,0.866,0.0,0.167,1.0,0.567,Medium
11,easy,What is first strike and how does it
work?,0.5,1.0,0.29,0.754,0.25,0.75,1.0,0.649,Medium
12,easy,Is Black Lotus legal in
Vintage?,1.0,1.0,0.8,0.866,0.5,0.0,1.0,0.738,High
13,easy,Can I play Counterspell in
Standard?,0.5,1.0,0.6,0.958,0.667,0.2,1.0,0.704,High
```

```
14,medium,What 2-mana blue counterspell is the most famous?,0.0,0.333,0.0,0.0,0.0,0.0,0.0,0.048,Low
15,medium,What red creatures cost 3 red mana (no colorless m...,0.0,0.889,0.0,0.928,0.0,0.0,0.0,0.26,Low
16,hard,If I control a creature with lifelink and deathtou...,0.667,0.0,0.62,0.931,0.143,0.0,1.0,0.48,Low
17,medium,Can I respond to someone casting a spell?,0.5,0.333,0.67,0.866,0.0,0.0,1.0,0.481,Low
18,hard,What happens if two players would simultaneously l...,1.0,0.5,0.0,0.934,0.5,0.0,1.0,0.562,Medium
```

What conclusions can you draw about the performance and effectiveness of your pipeline with this information?

Based on these evaluation results, we can see that the Stack Sage agent performs quite well on straightforward card lookup and basic rules questions (scoring High on 7 out of 18 questions), but struggles significantly with more complex scenarios that require specific tool usage or multi-step reasoning. The agent's verification layer is working as intended - it's correctly rejecting questions where it lacks sufficient context (like Q14 and Q15) rather than hallucinating answers, which explains the low scores for those questions but actually represents improved safety. The biggest opportunity for improvement lies in enhancing the search tools and entity extraction, particularly for questions that require finding specific cards by criteria or handling complex multi-card interactions.

Advanced Retrieval

Swap out base retriever with advnaced retrieval methods.

BM25 Added

Assessing Performance

How does the performance compare to your original RAG application? Test the new retrieval pipeline using the RAGAS frameworks to quantify any improvements. Provide results in a table.

```
question_id,difficulty,question_short,context_recall,faithfulness,factual_
correctness(mode=f1),answer_relevancy,context_entity_recall,noise_sensitiv
ity(mode=relevant),context_precision,overall_score,quality
1,easy,What does Lightning Bolt
do?,1.0,1.0,1.0,0.759,0.667,0.0,1.0,0.775,High
2,easy,What is the mana cost of
Counterspell?,1.0,1.0,1.0,1.0,0.333,0.0,1.0,0.762,High
3,easy,What type of card is Sol
Ring?,1.0,1.0,0.5,1.0,0.5,0.0,1.0,0.714,High
4,hard,"If my opponent controls Blood Artist, and one of
m...",1.0,1.0,0.86,0.926,1.0,0.333,1.0,0.874,High
5,hard,If my opponent has Doubling Season and I cast a
pl...,1.0,0.667,0.0,0.817,0.5,0.4,1.0,0.626,Medium
6,medium,How do Rest in Peace and Leyline of the Void
inter...,0.75,1.0,0.6,0.913,0.667,0.375,1.0,0.758,High
7,easy,If I cast Counterspell targeting my opponent's
Lig...,1.0,0.0,0.86,0.951,0.5,0.0,1.0,0.616,Medium
8,medium,Can I use Path to Exile on my own
creature?,1.0,1.0,0.86,0.976,0.25,0.25,1.0,0.762,High
9,medium,How does the stack
work?,1.0,0.667,0.92,0.708,0.167,0.222,1.0,0.669,Medium
10,medium,What are state-based
actions?,0.5,0.714,0.5,0.865,0.0,0.333,1.0,0.559,Medium
11,easy,What is first strike and how does it
work?,0.5,1.0,0.5,0.704,0.25,0.667,1.0,0.66,Medium
12,easy,Is Black Lotus legal in
Vintage?,1.0,1.0,0.8,0.843,0.5,0.0,1.0,0.735,High
13,easy,Can I play Counterspell in
Standard?,0.5,1.0,0.67,0.958,0.667,0.2,1.0,0.714,High
14,medium,What 2-mana blue counterspell is the most
famous?,0.0,0.0,0.5,0.945,0.0,0.0,0.0,0.206,Low
```

```
15,medium,What red creatures cost 3 red mana (no colorless
m...,0.0,0.875,0.0,0.88,0.0,0.0,0.0,0.251,Low
16,hard,If I control a creature with lifelink and
deathtou...,0.667,0.286,0.67,0.932,0.286,0.143,1.0,0.569,Medium
17,medium,Can I respond to someone casting a
spell?,1.0,0.667,0.67,0.919,0.0,0.0,1.0,0.608,Medium
18,hard,What happens if two players would simultaneously
l...,1.0,0.667,0.0,0.934,0.0,0.0,1.0,0.514,Medium
```

Results improved for some questions like the blood artist question, implying that the keyword research functionality of BM25 can prove useful, but i think the app is starting to have too many tools and is getting overwhelmed.

Articulate the changes that you expect to make to your app in the second half of the course. How will you improve your application?

I want to add better smart tool selection/get rid of unnecessary tools. I will improve this by having a query classification system, query analysis and routing, and generally improved tool orchestration.

Public Github Repo

A 5-minute (OR LESS) loom video of a live demo of your application that also describes the use case.

<https://www.loom.com/share/09a9f3998e694d8ebc906e6cbaa8624d?sid=e763769d-da19-4209-8554-b37dcebff1b5>

A written document addressing each deliverable and answering each question

This document!

All relevant code

https://github.com/ProfessorFess/Stack_Sage