

Regelfragen zu "Continuous Effects" bei Magic: The Gathering mithilfe von generativer KI beantworten

Wintersemester 2023/24

Inhaltsverzeichnis

1. Einleitung	3
1.1. Problemstellung	3
1.1.1. Exkurs: Grundlegende Regeln von <i>Magic: The Gathering</i>	3
1.1.2. Continuous Effects	4
1.1.3. Ziel	4
1.2. Stand der Forschung	5
1.2.1. Definition: ChatGPT	5
1.2.2. Definition: Prompt Engineering	5
1.2.3. Definition: Retrieval-Augmented Generation (RAG)	6
1.3. Technology Stack	6
1.3.1. Chat GPT-3.5-Turbo	6
1.3.2. LangChain 0.1.7	6
1.4. Daten	6
1.4.1. <i>Magic: The Gathering</i> Regelwerk	6
1.4.2. Kartentexte	7
2. Implementierung	8
2.1. Durchführung	8
2.2. Testfälle	9
2.2.1. <i>Urborg, Tomb of Yawgmoth</i> vs. <i>Blood Moon</i>	9
2.2.2. <i>Humility</i> vs. <i>Opalescence</i>	10
2.3. Prompt Template	10
2.4. Auswertung	11
3. Fazit	15
3.1. Ausblick	15
4. Literaturverzeichnis	16

1. Einleitung

Mit über 25.000 einzigartigen Karten und einem knapp 300-seitigen Regelwerk kann das Sammelkartenspiel *Magic: The Gathering* der Firma Wizard of the Coast konkurrenzlos als komplexestes Spiel der Welt gelten. Tatsächlich wurde 2019 bewiesen, dass *Magic: The Gathering* als einziges Brett- oder Kartenspiel mindestens die Komplexitätsstufe NP erreicht und sich aus einem turnierlegalem Deck eine Turingmaschine konstruieren lässt. [1]

Da die Autorin selbst Spielerin ist, entstand die Idee zu dieser Arbeit in der Regeldiskussion mit anderen Spieler:innen: In wiefern ist überhaupt möglich, dass Künstliche Intelligenz Regelfragen zu Spielsituationen beantwortet?

Im Folgenden wird die Fragestellung dieser Arbeit definiert, sowie eine Übersicht über relevante Spielmechaniken gegeben, um die untersuchten Testfälle zu erklären. Der Kontext der Arbeit wird in den aktuellen Forschungsstand eingeordnet, die Implementierung beschrieben, und abschließend die Ergebnisse ausgewertet. Für die Bearbeitung der gewählten Fragestellung ist Domänenwissen essentiell, ein Punkt, der im Bereich Machine Learning aktuell und von Bedeutung ist.

1.1. Problemstellung

Das Regelwerk von *Magic: The Gathering* ist nicht nur umfangreich, sondern auch in einer formalisierten Sprache geschrieben, und Begriffe erhalten innerhalb des Spiels neue oder erweiterte Bedeutungen.¹

Da die Regelfragen nur in den seltensten Fällen mit einem einfachen „ja“ oder „nein“ beantwortet werden können, konkretisierte sich die Idee, für die Beantwortung ein Large Language Model (LLM) zu verwenden. Eine Schwierigkeiten liegt darin, dass LLMs dazu neigen, Antworten zu halluzinieren, wenn sie diese nicht eindeutig wissen – unter Zuhilfenahme der Regeln aber jeder Streitfall geklärt werden kann, sofern sie richtig angewendet werden. Um dem LLM einen besseren Kontext zu vermitteln, muss ihm also das Regelwerk des Spiels zur Verfügung gestellt werden.

Die Forschungsfrage, mit der sich diese Arbeit auseinandersetzt, lautet: Ist es möglich, mithilfe eines Large Language Models Regelfragen zu „Continuous Effects“ des Kartenspiels *Magic: The Gathering* korrekt zu beantworten?

1.1.1. Exkurs: Grundlegende Regeln von *Magic: The Gathering*

Für das Verständnis dieser Arbeit ist es vonnöten, in aller Kürze die grundlegenden Regeln des Sammelkartenspiels zu erläutern: Zwei Spieler:innen wählen 60 Karten für ihr individuelles Deck aus, mit dem sie gegeneinander antreten. Um zu gewinnen, muss ein:e Spieler:in die 20 Lebenspunkte des Gegners auf 0 reduzieren. Dafür spielen sie Zaubersprüche, die entweder eine einmalige Wirkung haben, oder auf dem Spielfeld verbleiben (sog. Permanente); um diese Zaubersprüche wirken zu können, bezahlen sie mit der Ressource Mana, die von Ländern erzeugt wird.

¹Ein Beispiel wäre „Battlefield“: zwischen zwei Spieler:innen werden ausgespielte Karten abgelegt, die dort als sog. Permanente verbleiben. Der Regeltext liest sich wie folgt: 403.1. *Most of the area between the players represents the battlefield. The battlefield starts out empty. Permanents a player controls are normally kept in front of them on the battlefield, though there are some cases [...] when a permanent one player controls is kept closer to a different player. [...] 403.4. Whenever a permanent enters the battlefield, it becomes a new object and has no relationship to any previous permanent represented by the same card, except for the cases listed in rule 400.7. (This is also true for any objects entering any zone.)*



Abbildung 1: Aufbau einer *Magic: The Gathering* Karte. Die Textbox enthält den Regeltext und einen Flavour-Text.²

Die Karten werden beschrieben durch einen einzigartigen Namen und verfügen über Manakosten, einen Typ³, Fähigkeiten sowie bei Kreaturen Stärke und Widerstandskraft. Besonders die Fähigkeiten sind relevant, denn es gilt: Der Text einer Karte überschreibt die Regeln des Spiels.⁴

1.1.2. Continuous Effects

Die offiziellen Regeln definieren Continuous Effects wie folgt: *A continuous effect modifies characteristics of objects, modifies control of objects, or affects players or the rules of the game, for a fixed or indefinite period.* [2]

Der Großteil dieser Effekte und deren Auswirkungen auf die Spielfeldsituation kann trivial angewandt werden.⁵ In dieser Arbeit sollen komplexere Fälle untersucht werden, in denen die Continuous Effects von zwei Karten sich gegenseitig widersprechen. Die Entscheidung, welcher der Effekte angewandt wird, wird nach verschiedenen Kriterien getroffen:

- Die Reihenfolge der sogenannten Layer
- Einem Timestamp-System
- Einem Dependency-System

Die zwei gewählten Testfälle werden ausführlich in Abschnitt 2.2 beschrieben.

1.1.3. Ziel

Als Ergebnis des Projekts soll ein Jupyter Notebook vorliegen, in dem mit einer Abfrage an den Chatbot ChatGPT der Firma OpenAI eine Konflikt zwischen zwei sich widersprechenden Continuous Effects auf zwei Karten des Spiels *Magic: The Gathering* aufgelöst wird. Dazu soll ein Prompt Template erstellt

²Bildnachweis: <https://www.scouts.org.uk/activities/planeswalker-post/>, Zugriffen 13. März 2024

³Typen sind Kreaturen, Artefakte, Verzauberungen, Hexereien, Spontanzauber, Planeswalker, Schlachten sowie Länder.

⁴Eine grundlegende Regel des Spiels besagt: Ein:e Spieler:in zieht zu Beginn des Spielzuges eine Karte. Eine gespielte Karte erlaubt aber über ihren Kartentext, weitere Karten im Zug ziehen.

⁵Beispiel: Die Karte „Languish“: *All creatures get -4/-4 until end of turn.* Für eine definierte Zeit erhalten alle Kreaturen Auswirkungen auf ihre Stärke und Widerstandsfähigkeit.

werden, der für beliebige Karten einsetzbar ist. Im Rahmen des Projektes soll dieser Prompt für zwei Testfälle das korrekte Ergebnis liefern.

1.2. Stand der Forschung

Spätestens mit der Veröffentlichung und kostenfreien Nutzung von OpenAIs ChatGPT-3 im November 2022 erlebten die Sprachmodelle eine breite Aufmerksamkeit für eine Textgenerierung, die der menschlichen nahek kommt.

Ein bekanntes Problem großer Sprachmodelle ist die Halluzination bzw. Konfabulation. Hierbei werden in der Textgenerierung Antworten formuliert, die an sich zwar korrekt klingen, aber faktisch falsch sind; teilweise werden sogar Quellen erfunden. [3]

Weitere Probleme sind, dass die Daten, auf die sich Sprachmodelle beziehen, teilweise veraltet sind. Zudem werden häufig keine Quellen angegeben oder können nicht angegeben werden, sodass nicht nachvollzogen werden kann, wie das Ergebnis entstanden ist und ob es korrekt ist. Ein Lösungsansatz bietet hier Retrieval-Augmented Generation (RAG); hierbei kann dem Sprachmodell aktueller Kontext geliefert werden, um die gestellte Aufgabe zu bearbeiten. Die Alternative ist das Finetuning des Modells, das aber in dieser Arbeit nicht weiter betrachtet wird.

Im Zuge dieser Projektarbeit wird sich mit den aktuellen Themen der LLMs und RAG auseinandergesetzt, und versucht, eine Möglichkeit zu finden, mittels dem Sprachmodell ChatGPT-3.5 konkrete Regelfragen einer formalisierten Sprache eines komplexen Kartenspiels zu beantworten.

Aufgrund der Datenmenge und der Popularität des Kartenspiels gibt es einige Arbeiten im Bereich Machine Learning und Data Science mit *Magic: The Gathering* als Thema, jedoch geht es dort eher um Vorhersage von Kartenpreisen,⁶ Kartenfähigkeiten⁷ oder Zuordnung von Illustrationen.⁸ Hervorzuheben ist das Projekt „MTG-GPT“⁹, ein Finetuning von ChatGPT-3.5, das ebenfalls in der Lage ist, Regelfragen zu beantworten. Hierbei wird explizit RAG ausgeschlossen, und keine Karten- und Deckinformationen nach 2021 sind enthalten. Es ist unklar, ob durch das Finetuning mithilfe von „a set of Magic-specific questions“¹⁰ nur die Regelfragen der betreffenden Sonderfälle auswendig gelernt wurden, da es sich hierbei um häufig gestellte Fragen in Foren etc. handelt.

1.2.1. Definition: ChatGPT

ChatGPT ist ein Sprachmodell mit Generative Pre-trained Transformer (GPT) Architektur und zählt zu den Large Language Models. Es basiert auf der GPT-3.5 Architektur mit 6,7 Milliarden Parametern. Hervorzuheben sind die vielfältige Anwendungsmöglichkeiten im Bereich der Natural Language Processing (NLP), vor allem Textverständnis, Textgenerierung und Übersetzungen. [4] Die Trainingsdaten gewähren dem Modell nur ein Verständnis über Vorgänge bis zum Januar 2022.¹¹

1.2.2. Definition: Prompt Engineering

Das Konzept Prompt Engineering beschreibt im Bereich der Künstlichen Intelligenz die Eingabe einer Aufgabe in Textform an z. B. ein LLM. Diese verwendeten Text Prompts ermöglichen es, die Ausgaben des LLMs zu verbessern. Es gibt verschiedene Strategien wie das Gliederung in Teilaufgaben, Länge und Stil der Ausgabe und detaillierte Beschreibungen.¹²

⁶<https://is.muni.cz/th/jk7zd/PredictingThePriceOfMTGCards.pdf>, Zugriffen 11. März 2024

⁷https://cs229.stanford.edu/proj2015/187_report.pdf

⁸<https://arxiv.org/pdf/1810.03744.pdf>, Zugriffen 11. März 2024

⁹<http://mtg-gpt.com/>, Zugriffen 11. März 2024

¹⁰<https://medium.com/@zkiihne/introducing-mtg-gpt-c4ea7eb33582>, Zugriffen 11. März 2024

¹¹Stand März 2024

¹²<https://platform.openai.com/docs/guides/prompt-engineering>

1.2.3. Definition: Retrieval-Augmented Generation (RAG)

Um das Wissen außerhalb der Trainingsdaten von LLMs zu erweitern, wird Retrieval-Augmented Generation verwendet; statt Finetunings des Sprachmodells werden letztlich als Vektordatenbank gespeichert und können bei einem Task vom Sprachmodell als Kontextinformation verwendet werden. [5] In dieser Projektarbeit bietet das aktuelle Regelwerk von *Magic: The Gathering* den Kontext.

1.3. Technology Stack

Das Projekt wird mit Python 3 in einem Jupyter Notebook umgesetzt und ist auf Github einsehbar.¹³ Jupyter Notebook wurde gewählt, um die Änderungen am Prompt schnell zu testen und sowie die Textausgaben der Karteninformationen sowie Antworten des Chatmodells sofort im Notebook einzusehen.

Für die Paketverwaltung wird pip genutzt, als Virtual Environment virtualenv.

1.3.1. Chat GPT-3.5-Turbo

Das für das Projekt verwendete LLM ist ChatGPT-3.5-Turbo (im Folgenden kurz ChatGPT). Zu Beginn des Projekts im Wintersemester 2023/24 war ChatGPT und der API-Zugriff preiswerter im Vergleich zu GPT-4 und GPT-4-Turbo und wurde für die Anforderungen und den Umfang des Projektes als ausreichend eingeschätzt.¹⁴

1.3.2. LangChain 0.1.7

LangChain ist ein Framework¹⁵ für die Anwendungsentwicklung mit LLMs. In diesem Projekt wird es genutzt, um in einer Chain einen umfangreicheren Prompt sowie das gesamte Regelwerk von *Magic: The Gathering* als RAG zu übergeben.

Für die Generierung der Chains nutzt das Framework die LangChain Expression Language (LCEL); Chains sind eine Abfolge von Abfragen u. a. an ein LLM und werden hier genutzt, um die Informationen spezifischer Karten sowie das Regelwerk als Kontext an ChatGPT zu übergeben.

RAG wird als Alternative zum Finetuning des Modells genutzt, und um die maximal Token-Menge von 4097 bei ChatGPT-3.5 zu umgehen.

1.4. Daten

Das Regelwerk, die Karteninformationen sowie der Prompt werden auf Englisch verwendet. Das ausführliche Regelwerk liegt nur in englischer Sprache vor; desweiteren ist durch die formalisierte Regelsprache und das Verwenden der englischsprachigen Version von ChatGPT eine höhere Präzision zu erwarten.

1.4.1. Magic: The Gathering Regelwerk

Das offizielle Regelwerk mit dem Stand vom 2. Februar 2024 ist der Seite von Wizard of the Coast entnommen. Es umfasst 147.371 Wörter, was ungefähr 110.530 Token¹⁶ entspricht. Das Regelwerk ist als .txt im Projekt hinterlegt und wird ChatGPT als Vectorstore übergeben.

¹³https://github.com/Dornenhexe/project_mtg

¹⁴GPT-3.5-Turbo mit \$0,50 pro 1 Million Input Token gegenüber GPT-4-Turbo mit \$10,00 pro 1 Million Input Token. vgl. <https://openai.com/pricing>

¹⁵https://python.langchain.com/docs/get_started/introduction

¹⁶Laut <https://platform.openai.com/tokenizer> entspricht ein Token im Durchschnitt etwa einem dreiviertel Wort in der englischen Sprache.

1.4.2. Kartentexte

Alle relevanten Karteninformationen werden von der Seite Scryfall¹⁷ abgerufen. Die Karten-Daten von Scryfall sind in der Regel aktueller, von Fehlern bereinigt und per API abrufbar im Vergleich zur offiziellen Datenbank Gatherer.¹⁸

Die Informationen werden als .json gespeichert und zuvor von allen nicht benötigten Schlüsselwörtern befreit. Erhalten bleiben nur die Schlüssel, die für die Beantwortung von Regelfragen von Bedeutung sind: Name, Manakosten und umgewandelte Manakosten („cmc“), Farbe, Typ, Oracle-Text, produziertes Mana, Schlüsselwörter und bei Kreaturen Stärke/Widerstandsfähigkeit. Der Oracle-Text ist der Regeltext der Karte, vereinheitlicht für alle Versionen nach ggf. stattgefundenen Errata.

Zusätzlich werden für die Karten über ihre *id* noch zusätzliche Regelinformationen als *comments* abgerufen; diese enthalten offizielle Anmerkungen und Regelerklärungen. Gemeinsam mit den Schlüsselwörtern werden sie in einer Variable an den Prompt übergeben.

Bereinigtes JSON-File mit den entsprechenden Keys und zusätzlichen Regelinformationen für *Urborg, Tomb of Yawgmoth*:

```
{
  'id': '9e1a9e38-6ffc-490f-b0be-23ba4e8204c6',
  'name': 'Urborg, Tomb of Yawgmoth',
  'mana_cost': '',
  'cmc': 0.0,
  'type_line': 'Legendary Land',
  'oracle_text': 'Each land is a Swamp in addition to its other land types.',
  'colors': [],
  'keywords': [],
  'produced_mana': ['B'],
  'comments':
    ["Urborg, Tomb of Yawgmoth isn't a Swamp while it's not on the battlefield.",
     "Land cards not on the battlefield aren't Swamps while Urborg is on the battlefield.",
     'Urborg\'s ability causes each land on the battlefield to have the land type Swamp. Any land that\'s a Swamp has the ability "{T}: Add {B}." Nothing else changes about those lands, including their names, other subtypes, other abilities, and whether they\'re legendary, basic, or snow.',
     "If an effect such as that of Magus of the Moon causes Urborg to lose its abilities by setting it to a basic land type not in addition to its other types, it won't turn lands into Swamps, no matter in what order those effects started to apply."]
}
```

¹⁷<https://scryfall.com/>

¹⁸<https://gatherer.wizards.com/Pages/Default.aspx>

2. Implementierung

2.1. Durchführung

Im Vorfeld wurde das über das Web-Interface von ChatGPT getestet, in wie weit ein Verständnis für die Regeln von *Magic: The Gathering* vorhanden ist. Die Antworten lasen sich wie vom Chatbot gewohnt überzeugend, waren aber faktisch falsch, und konnten nur durch mehrfaches Nachfragen und Berichtigen ein akzeptables Ergebnis hervorbringen. Zudem entsprachen die zitierten Nummern des Regelwerks nicht dem des aktuellen.

Daraus ließen sich zwei grundlegende Ideen ableiten: Ein Prompt sollte mehrere Schritte enthalten, um die komplexe Aufgabe in einfacherer Unterschritte zu unterteilen. Des weiteren ist das Regelwissen des Models veraltet, und das neue Regelwerk vom Stand Februar 2024 muss als Kontext zur Beantwortung mit übergeben werden.

Beide Herausforderungen können mit dem Framework LangChain angegangen werden: Hier können umfangreiche Prompt Templates verfasst werden, die in mehrere Schritte gegliedert sind, eine sog. Chain of Thoughts. Den entsprechenden Kontext für das LLM liefert das Regelwerk mittels RAG.

Zuerst wurde die ChatGPT-API eingerichtet und erste Testanfragen mit einem kurzen Prompt versendet; ChatGPT hatte in diesem Fall Kenntnis von den Karten und ihrem Regeltext. Um beliebige Karten im Prompt zu verwenden und die für eine Regelentscheidung relevanten Informationen zu erhalten, wurden diese mithilfe der *requests* Library per HTTP GET bei scryfall.com erfragt. Das erhaltene JSON wurde wie in Abschnitt 1.4.2 beschrieben von ungenutzten Informationen bereinigt und gespeichert.

Um ChatGPT den Kontext zu liefern, wurde als Alternative zu einem Model Finetuning die Möglichkeit der Retrieved Augmented Generation gewählt. Hierfür bietet das Framework LangChain entsprechende Werkzeuge: Einen einfachen Loader, um das als Textdatei gespeicherte Regelwerk zu laden. Zu Beginn wurde mit einer gekürzten Version des Regelwerkes getestet, das nur die Abschnitte zu den Continuous Effects enthält. Folgend muss der Text in kleinere Chunks geteilt werden; hierfür wurde einer rekursiver Splitter gewählt, der nach Zeichen trennt. Zuerst wird der Text bei Absätzen geteilt, dann bei Satztrennzeichen und schließlich anhand von Leerzeichen in Wörter. Durch diesen Splitter bleiben semantisch zusammengehörige Textteile möglichst lange beieinander.

Die Textteile müssen im Anschluss in einem Vector Store gespeichert werden, damit dort Ähnlichkeitsabfragen möglich werden; d. h. dass es dem LLM möglich ist, Antworten auf die gestellte Frage anhand von semantischer Beziehung zu finden. Die Darstellung der Textteile erfolgt als Vektor und durch diese Darstellung bzw. Embedding wird die semantische Bedeutung eines Worts bewahrt. Als Vector Store wurde die opensource Datenbank Chroma gewählt, sowie die OpenAI-Embeddings für ChatGPT für die Embeddings. Das Ergebnis wird als Retriever gespeichert.

Die sog. Chain, mit der dem LLM der Kontext, Prompt Template und die eigentliche Frage übergeben wird, sieht in der LangChain Expression Language (LCEL) wie folgt aus:

```
chain = ({ "rules_db": retriever_rules, "question": RunnablePassthrough() } |  
prompt_template | model | StrOutputParser())
```

```
chain.invoke("What is the final board state of two cards played with continuous effects  
from Magic: the Gathering?")
```

Die Antworten des Chatmodels erfolgen als message type, weshalb sie zum String umgewandelt werden. Die Antworten werden zum Vergleich und zur Evaluation mit Datum und Uhrzeit als Textdatei gespeichert und können vollständig im Github-Projekt eingesehen werden.

Da die Beantwortung zu den Regelfragen nicht vollständig korrekt waren – entweder wurde nur die falsche Begründung gegeben oder sogar vollkommen eine Entscheidung getroffen, die den Regeln wi-

derspricht – wurde fortlaufend der Prompt überarbeitet. Eine korrekte Ausgabe eines Testfalles führte nicht auch zu einem korrekten Ergebnis für den zweiten Fall.

Im weiteren Verlauf wurde nun das gesamte Regelwerk statt der gekürzten Fassung als Kontext übergeben.

Um weiteren Kontext zu liefern, wurden nun auch die erweiterten Regeln von scryfall.com in den Karteninformationen gespeichert. Oftmals erhalten sie eine offizielle Spezifizierung der Regeltexte und teilweise auch Erläuterungen von Grenzfällen, allerdings kein konkretes Beispiel für die gewählten Testfälle.¹⁹

Eine Auswertung der Antworten erfolgt unter Abschnitt 2.4.

2.2. Testfälle

Die meisten Continuous Effects sind im Spielverlauf einfach anzuwenden und sind additiv.²⁰ Für die Projektarbeit werden zwei Testfälle untersucht, deren Continuous Effects zu Konflikten führen, die auch für Spieler:innen oftmals nicht nur mithilfe des Kartentextes zu klären sind.

2.2.1. Urborg, Tomb of Yawgmoth vs. Blood Moon



Abbildung 2: Urborg, Tomb of Yawgmoth und Blood Moon

Konflikt: In diesem Fall kommt es zum Konflikt, denn *Urborg, Tomb of Yawgmoth* ist kein Standardland, denn diese besitzen nur den Typ „Basic Land“ sowie die Fähigkeit, für ein Mana ihrer Farbe getappt zu werden. *Blood Moon* ändert aber die Fähigkeiten aller Nicht-Standardländer. Erhält *Blood Moon* aber zusätzlich den Typ *Swamp* oder überschreibt es zuvor die Fähigkeit von *Urborg, Tomb of Yawgmoth*? Die Continuous Effects beider Karten geschehen in derselben Layer:

Layer 4: Type-changing effects are applied. These include effects that change an object's card type, subtype, and/or supertype. [2]

¹⁹So wird zwar der Fall *Urborg, Tomb of Yawgmoth* und *Magus of the Moon* untersucht, der zwar denselben Oracle-Text hat wie *Blood Moon*, dieser aber in der Regelerweiterung nicht genannt wird. Für *Humility* und *Opalescence* werden Beispiele erklärt, in denen zwei *Opalescence* im Spiel sind.

²⁰Sowohl *Bad Moon* erhöht mit *Black creatures get +1/+1* Stärke und Widerstandsfähigkeit von Kreaturen, und auch *Intangible Virtue* verändert die Werte von Schwarzen Kreaturenspielsteinen mit *Creature tokens you control get +1/+1 and have vigilance*, so werden die Werte addiert.

Erwartete Ausgabe: Alle Nicht-Standardländer werden zu Gebirgen, also auch *Urborg*, *Tomb of Yawgmoth* und verliert seine Fähigkeit. In diesem Fall spielt es keine Rolle, welche Karte zuerst gespielt wurde: *Blood moon* verwandelt *Urborg*, *Tomb of Yawgmoth* in ein Gebirge, noch bevor es das Spielfeld betritt. Somit ist der Effekt einer Karte von der anderen abhängig.

Die Entscheidung wird anhand der folgenden Regel getroffen:

613.8. *Within a layer or sublayer, determining which order effects are applied in is sometimes done using a dependency system. If a dependency exists, it will override the timestamp system.* [2]

2.2.2. Humility vs. Opalescence



Abbildung 3: *Humility* und *Opalescence*

Konflikt: Wenn *Humility* gespielt wird, werden alle Kreaturen zu Kreaturen mit Stärke und Widerstandsfähigkeit von 1/1. Aber *Opalescence* besagt, dass jede Verzauberung (in diesem Fall: *Opalescence* und *Humility*) Kreaturen mit Stärke und Widerstandsfähigkeit von 4/4 (aufgrund ihrer Umgewandelten Manakosten in Höhe von vier) sein werden.

Das Ergebnis hängt davon ab, welche Karte zuerst gespielt wird, deshalb wird dieser Fall auch jeweils in umgekehrter Reihenfolge getestet.

Erwartete Ausgabe: Wenn erst *Humility*, dann *Opalescence* gespielt wird, haben alle Kreaturen eine Stärke und Widerstandsfähigkeit von 4/4. Wird zuerst *Opalescence* gespielt und anschließend *Humility*, haben alle Kreaturen eine Stärke und Widerstandsfähigkeit von 1/1.

Dieser Konflikt wird durch die Timestamp-Regel gelöst:

613.7. *Within a layer or sublayer, determining which order effects are applied in is usually done using a timestamp system. An effect with an earlier timestamp is applied before an effect with a later timestamp.* [2]

2.3. Prompt Template

Das Prompt Template soll allgemein gültig sein für Regelfragen betreffend Continuous Effects. Um das Ergebnis der Ausgabe zu verbessern, muss der Prompt in mehreren Iterationen angepasst werden und ständig getestet werden, da weder vorhersehbar noch deterministisch ist, welche Änderung am Prompt welche Ausgabe mit sich zieht.

Der Aufbau des Prompts orientiert sich an den Richtlinien von OpenAI zum Prompt Engineering²¹ und ist wie folgt gegliedert:

1. Die Vergabe einer Rolle
2. Details zur Fragestellung
3. Kontext im Rahmen der Aufgabenstellung und wie Daten zu behandeln sind
4. Daten als Kontext via RAG geben
5. Teilung der Aufgabe in nummerierte Unteraufgaben

Das finale Prompt Template ist folgendes:

You are judge (aka a referee) in a game of Magic: the Gathering. You will get a {question} regarding {card1} and {card2} and give an answer.
Use the {rules_db} as context for your decision. {card1} and {card2} each contain information of a single Magic: The Gathering card.

The following things are true:

- {card1} and {card2} both have continuous effects.
- Consider using the "comments" from {card1} and {card2}.
- The "type_line" of the card could be relevant for ruling decision.
- {card1} is played first and has the earlier timestamp, {card2} is played second and has the later timestamp. This may be relevant for the decision, but shouldn't influence your decision which rule to apply.
- Only one rule determine the outcome.

You will follow these steps to make a decision:

1. Give the oracle texts from both cards as well the comments.
2. Determine the layer on which the effects of the cards are applied. Layers are defined in 613.1 in the {rules_db}.
3. Decide which rule you think is fitting: Timestamp and Layers (613.1 to 613.7), Dependency (613.8) or Overriding of effects (613.9). Consider the card texts and comments from both cards for your decision.
4. Summarize the rule of your choosing.
5. Give your judging regarding the chosen rule: The continuous effect of which card is applied in game: {card1} or {card2}? Quote the effects.
6. Explain why you decided this way.

Die Anpassungen an das Prompt Template können im Projekt unter https://github.com/Dornenhexe/project_mtg im Ordner „data“ in der Datei „old_prompts.txt“ angesehen werden.

2.4. Auswertung

Erste Teilerfolge stellten sich früh ein: Die Karten wurden korrekt erkannt, die Aufgabestellung wurde verstanden und eine Antwort gegeben. Die ersten Antworten waren jedoch nicht korrekt und es kam zu LLM-typischen Konfabulationen, bei denen Regelentscheidung und geforderte Erklärung der Regeln sich widersprachen.²²

Hierdurch wurden die Schwachstellen des frühen Prompt Templates deutlich:

You will get two cards from the cardgame Magic: the Gathering. {card1} and {card2} both have continuous effects. You need to determine which effect will have which impact on the game.

You can assume that {card1} is played first, therefor has the earlier timestamp.

²¹<https://platform.openai.com/docs/guides/prompt-engineering>

²²vgl. z. B. Antwort 2024-02-28 11-33-42.txt, in der sich auf die Regel 613.9 „Overriding of effects“ bezogen wird, aber mit der der Timestamp-Regel 613.7 argumentiert wird.

Use the {rules_db} as context for your decision.
You will get a {question} about two cards.

You will follow these steps:

1. Give the oracle texts from both cards.
2. Give your judging: The continuous effect of which card is applied in game: {card1} or {card2}?
3. Explain why you decided this way.
4. Quote the rule you used for your decision.

Es gab zu wenig Erklärung der Karten und Einordnung der Regeln, außerdem war vermutlich der starke Hinweis auf den Timestamp verantwortlich dafür, dass jede Regelentscheidung mit der Timestamp-Rule begründet wurde. In der frühen Phase des Projekts wurde nur das gekürzte Regelwerk übergeben, das nur den Ausschnitt über die Continuous Effects enthält.

In iterativen Anpassungen wurden verschiedene Parts des Prompts angepasst; hierbei gelang meist eine korrekte Aussage für ein oder zwei der drei Fälle. Bereits kleinste Änderungen hatten zur Folge, dass ChatGPT für die Antwort eine andere Regel für die Beantwortung der Regelfrage wählte. Als besonders hilfreich erwies sich das einschränken der Formulierung bzgl. des Timestamps („may be relevant“) und die Option zwischen den drei Regeln zu geben (vgl. finalen Prompt unter Abschnitt 2.3). Wird allerdings Punkt 3 „Timestamp and Layers (613.1 to 613.7)“ reduziert auf „Timestamp (613.7)“, kommt es in allen Testfällen zu falschen Begründungen.

Mit der zusätzlichen Übergabe von den weiteren Karteninformationen und dem gesamten Regelwerk konnten korrekte Antworten für alle drei Fälle generiert werden. Allerdings war dies bisher nicht mit dem finalen Prompt möglich; maximal zwei von drei Fällen werden korrekt behandelt.

Der Fall *Urborg, Tomb of Yawgmoth* vs. *Blood Moon* in der Antwort „case1_2024-03-14 13-17-15.txt“ gibt die entsprechende Regel an und trifft die Regelentscheidung wie unter Abschnitt 2.2 definiert:

1. Oracle texts and comments:

- *Urborg, Tomb of Yawgmoth*: Each land is a Swamp in addition to its other land types. Comments include details about how Urborg's ability works on the battlefield and interactions with other effects.
- *Blood Moon*: Nonbasic lands are Mountains. Comments explain the implications of this effect on nonbasic lands.

2. Determine the layer:

The effects of both cards are continuous effects, and they both affect the type of lands on the battlefield. According to the rules, effects that change the characteristics of objects are applied in layer 4 (type-changing effects).

3. Decide on the rule:

In this scenario, we will apply the rule of Dependency (613.8) because the effects of both cards are dependent on each other. *Urborg, Tomb of Yawgmoth*'s ability to make lands Swamps is dependent on the existing land types, and *Blood Moon*'s effect to turn nonbasic lands into Mountains is dependent on the lands being nonbasic.

4. Rule summary:

Dependency (613.8) states that when two continuous effects are dependent on each other, the dependent effect is applied first.

5. Judging:

The continuous effect of *Blood Moon* is applied in the game. The lands on the battlefield will be turned into Mountains by *Blood Moon*'s effect.

6. Explanation:

We chose the rule of Dependency because the effects of both Urborg, Tomb of Yawgmoth and Blood Moon are dependent on each other. Blood Moon's effect to change nonbasic lands into Mountains is dependent on the lands being nonbasic, and Urborg's ability to make lands Swamps is dependent on the existing land types. Therefore, Blood Moon's effect takes precedence in this scenario.

Zusätzlich wird einer der beiden Fälle *Opalescence* vs. *Humility* bzw. *Humility* vs. *Opalescence* zwar korrekt beantwortet, allerdings falsch begründet (nicht die Reihenfolge der Layer ist entscheidend, sondern welche Karte zuerst gespielt wird). Beachtenswert ist hierbei, dass es sich um dieselben Karten handelt, bei der dieselbe Regel angewendet wird, hier nur die Reihenfolge relevant ist, in der die Karten gespielt werden.

Für das finale Prompt Template erfolgt für den Fall *Opalescence* vs. *Humility* folgende Regelbeurteilung (case2_2024-03-14 13-17-53.txt):

1. **Opalescence**:

- Oracle Text: Each other non-Aura enchantment is a creature in addition to its other types and has base power and base toughness each equal to its mana value.
- Comments: The interaction between Humility and Opalescence is explained, detailing the effects on different layers depending on the timestamp order.

2. **Humility**:

- Oracle Text: All creatures lose all abilities and have base power and toughness 1/1.
- Comments: The interaction between Humility and Opalescence is described, including the layering of effects and the impact on power and toughness.

3. The effects of both cards are applied in different layers. Opalescence's effect is applied in Layer 4, while Humility's effect is applied in Layer 6.

4. **Rule Choice: Layers (613.1 to 613.7)**

- This rule governs the order in which continuous effects are applied based on the layers defined in the game.

5. **Judging Decision**:

- According to the Layers rule, the effect of **Humility** takes precedence over **Opalescence** because it is applied in a higher layer (Layer 6 vs. Layer 4).
- Therefore, in the final board state, all creatures will be 1/1 with no abilities, as dictated by Humility's effect.

6. **Explanation**:

- The decision to apply the Layers rule is based on the understanding of how continuous effects interact in Magic: The Gathering. By following the layering system defined in the rules, we can determine the order in which effects are applied and resolve any conflicts between them.
- In this case, Humility's effect of reducing all creatures to 1/1 with no abilities overrides Opalescence's effect of turning enchantments into creatures with power and toughness equal to their mana value.
- This ruling ensures consistency and fairness in gameplay by following the established rules of the game.

Dieselbe Begründung wird auch für den umgekehrten Fall (vgl. case2_2024-03-14 13-22-39.txt) gegeben, dadurch wiederum ist die Entscheidung aber falsch. Mit einem verstärkten Hinweis auf die Timestamp-Regel (vgl. Prompt 4 in old_prompts) ist es möglich, die hier eine korrekte Antwort zu

erhalten (vgl. case2_2024-03-04 16-16-26.txt), dann wird aber ebenfalls für den ersten Testfall die Timestamp-Regel angewandt.

Ob eine Antwort korrekt gegeben wurde, hing z. T. stark von geringen Änderungen im Prompt Template ab. Vereinzelt kam es vor, dass die Ausgaben sich mit demselben Prompt nicht replizieren ließen, obwohl die Temperatur des ChatGPT Modells auf 0 gesetzt war, was die Ausgabe weniger kreativ, aber präziser werden lassen soll. Diese Unzuverlässigkeiten im Testen erschwerten die Erstellung eines allgemeinführenden Prompt Templates.

Korrekte oder zum Teil korrekte Antworten wurden im Ordner „answers“ nach Datum, Uhrzeit und Testfall benannt gespeichert und können dort eingesehen werden. Nicht jede Antwort wurde gespeichert, da sie sich oftmals wiederholten oder auch beim Testen einer Veränderung im Prompt wieder ein falsches Ergebnis lieferten.

3. Fazit

Im Rahmen dieser Projektarbeit ist es nicht gelungen, alle definierten Testfälle mittels generativer KI zuverlässig zu beantworten. Das finale Prompt Template liefert eingeschränkt auf einen Testfall das erwartete Ergebnis, aber es ist kein allgemeingültiges Prompt Template für alle Fälle, wie es Ziel der Aufgabe war. Es ist aber nicht ausgeschlossen, dass es mit anderen Formulierungen im Prompt möglich ist oder einem anders strukturiertem Prompt Template.

Die Ergebnisse sind in der Hinsicht vielversprechend: Es werden zum Großteil korrekte Ausgaben geliefert, und auch die sehr konkreten Regelfragen des umfassenden Regelwerks von *Magic: The Gathering* werden richtig angewandt. Eine Annahme ist, dass neuere Modelle von OpenAI oder anderen Anbietern auch bessere, verlässlichere Ergebnisse erzielen würden.

Da aber die Antworten nicht deterministisch waren, sondern vereinzelt andere Ausgaben mit demselben Prompt erzeugt wurden, kann das Ergebnis der Projektarbeit zu diesem Stand nicht zuverlässig verwendet werden, um Regelfragen mithilfe von generativer KI zu beantworten.

RAG hat in dieser Projektarbeit das kosten- und zeitintensive Finetuning des ChatGPT-Modells ersetzt und den benötigten Kontext für das Chatmodell geliefert. Auch wenn das in Abschnitt 1.2 erwähnte Projekt „MTG-GPT“ die Regelfragen der Testfälle korrekt beantworten kann, ist ungewiss, mit welchen Daten das Finetuning erfolgte. Falls Foren und Blogeinträge verwendet wurden, ist nicht auszuschließen, dass hierbei auch Antworten zu Regelfragen gelernt wurden, die auf einer fehlerhaften Interpretation der Regeln beruhen.

Die Kosten des Projekts beliefen sich auf ca. \$1 für die Verwendung der ChatGPT-API.

3.1. Ausblick

Um das Prompt-Template zuverlässig für die allgemeine Beantwortung zu Regelfragen zu Continuous Effects zu formulieren, müssten weit aus mehr Fälle getestet werden. Diese müssen dann auch neben der Timestamp- und Dependency-Rule weitere Regeln betreffen. Eine weitere Option für ein optimiertes Prompt-Template wäre die Verwendung von Directional Stimulus Prompting (DSP), bei dem Hinweise für spezifische Tasks gegeben werden.

Um dies zu unterstützen könnten die Karten im Vorfeld vorsortiert werden: Denkbar wäre eine Einordnung in Layer, wie es im aktuellen Prompt schon geschieht, aber auch ein Filtern nach bestimmten Signalwörtern, die Rückschluss auf eine Regelanwendung geben. Dies könnte ebenfalls über ChatGPT und eine erweiterte Chain gelöst werden, vielleicht aber auch durch eine algorithmische Vorsortierung mit semantischer Filterung.

Ein weitere Option ist das Testen weiterer Chatbots; während noch zu Beginn des Projektes im September 2023 ChatGPT als beste Option galt, sind bereits ein halbes Jahr später im März 2024 weitere Auswahlmöglichkeiten erschienen, die in Betracht zu ziehen sind, wie Anthropic Claude 3.²³

Für eine Anwendung, die über die theoretische Betrachtung hinaus geht, kann das Jupyter Notebook durch ein Refactoring als Python-Anwendung umgesetzt werden. Diese könnte dann beispielsweise anderen Spieler:innen als (Web-)App zur Verfügung gestellt werden. Es ist zu erwarten, dass die Zielgruppe für diese Art von spezifischen Regelfragen eher klein ausfällt.

²³https://www.heise.de/news/Anthropic-Claude-3-schneidet-in-Tests-besser-ab-als-GPT-4-9645971.html?wt_mc=sm.red.ho.mastodon.mastodon.md_beitraege.md_beitraege, Zugriffen 11. März 2024

4. Literaturverzeichnis

- [1] A. Churchill, S. Biderman, und A. Herrick, „Magic: The Gathering is Turing Complete“. Zugegriffen: 29. Februar 2024. [Online]. Verfügbar unter: <https://arxiv.org/abs/1904.09828>
- [2] W. of the Coast, „Magic: The Gathering Comprehensive Rules“. Zugegriffen: 1. März 2024. [Online]. Verfügbar unter: <https://media.wizards.com/2023/downloads/MagicCompRules20230616.txt>
- [3] H. Alkaissi und S. Mcfarlane, „Artificial Hallucinations in ChatGPT: Implications in Scientific Writing“, *Cureus*, Bd. 15, S. , 2023, doi: [10.7759/cureus.35179](https://doi.org/10.7759/cureus.35179).
- [4] P. P. Ray, „ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope“, *Internet of Things and Cyber-Physical Systems*, Bd. 3, S. 121–154, 2023, doi: <https://doi.org/10.1016/j.iotcps.2023.04.003>.
- [5] P. Lewis u. a., „Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks“. Zugegriffen: 5. März 2024. [Online]. Verfügbar unter: <https://arxiv.org/abs/2005.11401>