DHBW
Duale Hochschule
Baden-Württemberg

Evaluierung von LLM-basiertem QA

Studienarbeit

Studiengang Informatik

Duale Hochschule Baden-Württemberg Stuttgart

Anton Seitz

Eingereicht am: 21.03.2025
Matrikelnummer, Kurs: 3626401, INF22B
Betreuer an der DHBW: Dr. Armin Roth

Zusammenfassung

Diese Arbeit untersucht systematisch die Antwortqualität von LLM-basierten QA-Systemen. Aufbauend auf den Erkenntnissen aus [1] und [2] wird ein neues Test-Environment entwickelt, in dem ein speziell definierter Textkorpus – beispielsweise aus judo-spezifischen Quellen und Wikipedia-Artikeln – sowie gezielte Variationen der Fragestellungen zur Evaluation herangezogen werden. Ziel ist es, mittels quantitativer Metriken (z. B. Genauigkeit, Präzision, Recall, F1-Score) und qualitativer Analysen die Leistungsfähigkeit der #acrplLLM hinsichtlich Korrektheit und Vollständigkeit der generierten Antworten zu bewerten. Die gewonnenen Erkenntnisse sollen die aktuellen Grenzen der Modelle aufzeigen und Ansatzpunkte für deren Optimierung in praktischen Anwendungen liefern.

Inhalt

1

Kurzbeschreibung der Arbeit

Diese Studienarbeit befasst sich mit der Evaluierung von Large Language Model (LLM)-basiertem Question Answering (QA). Dabei wird untersucht, wie LLMs in der Lage sind, natürlich formulierte Fragen basierend auf einem definierten Textkorpus automatisch zu beantworten. Der Textkorpus wird aus fachspezifischen Quellen (z. B. judo-spezifische Literatur, Wettkampfdaten) und ergänzend aus öffentlich zugänglichen Daten wie Wikipedia-Artikeln zusammengestellt. Die Ausgangssituation ist geprägt durch den rasanten Fortschritt im Bereich der künstlichen Intelligenz und Natural Language Processing (NLP). Trotz der beeindruckenden Leistungsfähigkeit weisen aktuelle Modelle oftmals Defizite in Bezug auf Faktenwissen und Detailgenauigkeit auf. Ziel dieser Arbeit ist es, die Antwortqualität – gemessen an Kriterien wie Korrektheit, Vollständigkeit und Relevanz – systematisch zu bewerten und potenzielle Einsatzbereiche zu identifizieren.

2

Einleitung

Dieses Kapitel führt in das Themengebiet ein und legt die Motivation sowie die Problemstellung dar.

2.1 Motivation

Führende LLMs werden im Zuge des aktuellen Hypes häufig als Alleskönner dargestellt. Wie in [1] ersichtlich, weisen diese Modelle jedoch oftmals ein mangelhaftes Faktenwissen auf und beantworten selbst bei bekannten Informationen natürlichsprachliche Fragen fehlerhaft. Sogar das beste getestete Modell, GPT-4, erreichte lediglich eine Korrektheitsrate von 40,3 %. Diese Diskrepanz zwischen den Erwartungen und der tatsächlichen Leistungsfähigkeit bildet die Motivation für diese Arbeit. Ziel ist es, die Grenzen der Leistungsfähigkeit von LLMs systematisch zu erforschen.

2.2 Zielsetzungen

Die Arbeit verfolgt folgende Ziele:

1. Entwicklung eines Test-Environments zur systematischen Evaluierung von LLM-basierten QA-Systemen.
2. Bewertung der Antwortqualität anhand quantitativer und qualitativer Metriken.
3. Identifikation von Einsatzbereichen, in denen LLM-basierte QA-Systeme einen Mehrwert bieten können.

3

Grundlagen und Definitionen

Im folgenden werden die Grundlagen und Funktionsweisen von LLMs sowie die Konzepte von QA erläutert.

3.1 LLMs

Dieses Kapitel beleuchtet die Grundlagen und Funktionsweisen von LLMs.

3.1.1 Architektur und Trainingsparadigmen

Es werden die zugrundeliegenden tiefen neuronalen Netzwerke, die Rolle umfangreicher Trainingsdatensätze und die Auswirkungen verschiedener Trainingsparadigmen erläutert.

3.1.2 Stärken und Limitationen

Die Leistungsfähigkeit der Modelle wird kritisch diskutiert – basierend auf Erkenntnissen aus [1] – wobei auch die bekannten Schwächen im Bereich des Faktenwissens thematisiert werden.

3.1.3 Stand der Forschung

Ein Überblick über aktuelle Entwicklungen und Forschungsarbeiten im Bereich LLMs bietet die Basis für diese Arbeit.

3.2 Question Answering

Hier wird das Konzept des QA umfassend dargestellt.

3.2.1 Modelltypen: Open Domain vs. Closed Domain

QA-Systeme können in zwei Kategorien unterteilt werden:

- Open Domain QA: Systeme, die mit breit gefächerten, öffentlichen Wissenskorpora (z. B. Wikipedia) arbeiten und allgemeine Fragen beantworten.
- Closed Domain QA: Systeme, die auf spezifische, thematisch eingeschränkte Korpora (z. B. judo-spezifische Literatur) fokussiert sind und detaillierte, fachspezifische Informationen liefern.

3.2.2 Typen von Question Answering Modellen

- Extractive Question Answering: Ein tief lernendes Modell, das eine Antwort liefert, wenn ein Textkorpus (also ein Kontext) gegeben ist. Das Modell „durchsucht“ die Dokumente, um die beste Antwort auf die Frage zu finden. Es funktioniert im Wesentlichen wie ein Suchwerkzeug.
- Open Generative Question Answering: Ein Modell, das auf Grundlage eines Kontextes Text generiert. Im Gegensatz zum extraktiven Modell muss die Antwort nicht wörtlich im Text stehen.
- Closed Generative Question Answering: Ein Modell, bei dem kein Kontext bereitgestellt wird und die Antwort vom Modell generiert wird.

3.2.3 Methoden des QA

Vorstellung der unterschiedlichen Ansätze, von retrievalbasierten Methoden – bei denen relevante Textpassagen aus einem Korpus extrahiert werden – bis hin zu generativen Ansätzen, bei denen Antworten synthetisch erstellt werden.

3.2.4 Implementierungspipelines

Erläuterung, wie QA-Systeme in der Praxis umgesetzt werden. Beispielsweise wird die Integration von BERT- bzw. RoBERTa-basierten Modellen anhand des in [2] vorgestellten Konzepts erläutert.

4

Metriken und Evaluationsmethoden

Dieses Kapitel beschreibt die Kennzahlen und Verfahren zur Bewertung der QA-Systeme.

4.1 Quantitative Kennzahlen

Detaillierte Beschreibung von Metriken wie Genauigkeit, Präzision, Recall und F1-Score zur objektiven Messung der Antwortqualität.

4.2 Qualitative Evaluationsansätze

Erörterung von Verfahren zur inhaltlichen Bewertung, beispielsweise durch Expertenreviews oder den Vergleich mit vorab definierten Referenzantworten.

4.3 Validierung der Ergebnisse

Methoden zur Überprüfung der Reproduzierbarkeit und Aussagekraft der experimentellen Daten werden vorgestellt.

5

Konzept

Das konzeptionelle Vorgehen bei der Entwicklung des Test-Environments wird hier erläutert.

5.1 Systemarchitektur

Beschreibung der geplanten Architektur, inklusive der Integration des ausgewählten LLM und der Anbindung an den definierten Textkorpus, der aus judo-spezifischen Quellen und ergänzend aus Wikipedia besteht.

5.2 Design des Test-Environments

Ausarbeitung der Strategien zur Generierung von Testfragen und der Festlegung von Referenzantworten. Dabei wird ein flexibles Framework entwickelt, das an verschiedene Evaluationsszenarien angepasst werden kann.

6

Realisierung

Dieses Kapitel beschreibt die praktische Umsetzung des Konzepts.

6.1 Implementierungsdetails

Die QA-Pipeline wird mithilfe der Huggingface-Transformers-Bibliothek implementiert. Dabei wird ein vortrainiertes Modell, wie z. B. **deepset/roberta-base-squad2**, genutzt. Der Aufbau der Pipeline orientiert sich an einem modifizierten Python-Skript, das an das in [2] gezeigte Colab-Beispiel angelehnt ist. Dieses Skript übernimmt unter anderem folgende Aufgaben:

- Laden des Modells und Tokenizers
- Übergabe von Frage- und Kontextdaten zur Inferenz
- Ausgabe der generierten Antwort samt zugehöriger Metriken

6.2 Software- und Hardwareumgebung

Beschreibung der eingesetzten technischen Ressourcen, wie Graphics Processing Unit (GPU)-gestützte Server oder Cloud-Services, die zur Beschleunigung der Inferenz und eventueller Experimente genutzt werden.

6.3 Test- und Validierungsszenarien

Darstellung der durchgeführten Tests, einschließlich der Variation von Fragestellungen (sowohl Open Domain als auch Closed Domain) und der kontinuierlichen Evaluierung der Zwischenergebnisse mittels der zuvor definierten Metriken.

6.4 Datenbeschaffung

Der Textkorpus wird durch die Aggregation verschiedener Quellen erstellt:

- Fachspezifische Literatur: Judo-Regelwerke, Trainingshandbücher und Wettkampfdaten.
- Öffentliche Quellen: Artikel und Einträge von Wikipedia, die durch Web-Scraping oder APIs bezogen werden können.
- Eigene Notizen: Ergänzende, vom Autor erstellte Daten, um den Korpus zu erweitern.

7

Evaluierung

In diesem Kapitel werden die experimentellen Ergebnisse analysiert und interpretiert.

7.1 Ergebnisauswertung

Systematische Auswertung der quantitativen und qualitativen Messergebnisse. Ergebnisse werden grafisch (z. B. in Diagrammen) dargestellt.

7.2 Diskussion der Resultate

Kritische Analyse der Resultate im Hinblick auf die definierten Zielsetzungen und identifizierten Limitationen der LLMs.

7.3 Vergleich mit bestehenden Ansätzen

Die erarbeiteten Ergebnisse werden mit den in der Literatur beschriebenen Ansätzen (insbesondere [1] und [2]) verglichen, um den Mehrwert des entwickelten Systems herauszustellen.

8

Zusammenfassung und Ausblick

Das abschließende Kapitel fasst die gewonnenen Erkenntnisse zusammen und gibt einen Ausblick auf zukünftige Entwicklungen.

8.1 Schlussfolgerungen

Zusammenfassung der wesentlichen Ergebnisse und Bewertung, inwiefern die definierten Ziele erreicht wurden.

8.2 Empfehlungen für zukünftige Arbeiten

Ableitung von Empfehlungen und potenziellen Erweiterungen für weiterführende Forschungen im Bereich LLM-basiertes QA.

8.3 Reflexion und Ausblick

Kritische Reflexion der Limitationen der durchgeführten Experimente und ein Ausblick auf zukünftige technologische Entwicklungen.

A Literatur

[1] K. Sun, Y. E. Xu, H. Zha, Y. Liu, und X. L. Dong, „Head-to-tail: how knowledgeable are large language models (LLMs)? AKA will LLMs replace knowledge graphs?“, *arXiv preprint arXiv:2308.10168*, 2023.

[2] „Question Answering with BERT“. 2023.

B Abkürzungen

API	Application Programming Interface
GPU	Graphics Processing Unit
HTTP	Hypertext Transfer Protocol
LLM	Large Language Model
NLP	Natural Language Processing
QA	Question Answering
REST	Representational State Transfer

C Glossar

Komponente Ein Architekturbaustein. Zusammengesetzte Komponenten bestehen aus weiteren Subkomponenten. Einfache Komponenten sind nicht weiter unterteilt.

Softwareschnittstelle Ein logischer Berührungspunkt in einem Softwaresystem: Sie ermöglicht und regelt den Austausch von Kommandos und Daten zwischen verschiedenen Prozessen und Komponenten.

Selbstständigkeitserklärung

Gemäß Ziffer 1.1.13 der Anlage 1 zu §§ 3, 4 und 5 der Studien- und Prüfungsordnung für die Bachelorstudiengänge im Studienbereich Technik der Dualen Hochschule Baden- Württemberg vom 29.09.2017. Ich versichere hiermit, dass ich meine Arbeit mit dem Thema:

Evaluierung von LLM-basiertem QA

selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Ich versichere zudem, dass alle eingereichten Fassungen übereinstimmen.

Stuttgart, 21.03.2025

Anton Seitz