Inhaltsverzeichnis

1 Einführung

Die Fortschritte in der Künstlichen Intelligenz, insbesondere im Bereich der Verarbeitung natürlicher Sprache (NLP), haben zu Durchbrüchen im "Question Answering"(QA) geführt. Traditionell wurde diese Aufgabe durch das Fine-Tuning von Transformer-Modellen wie DistilBERT und T5 für spezifische Datensätze gelöst. Mit der Einführung großer vortrainierter Sprachmodelle (LLMs) wie ChatGPT und LLAMA 2 ist es nun möglich, durch geschickte Prompt-Gestaltung Antworten basierend auf dem Kontext zu generieren, ohne ein Modell spezifisch für einen Datensatz zu trainieren.

2 Projektbeschreibung

2.1 Ziele der Studienarbeit

Das Ziel dieser Studienarbeit ist es, Fine-Tuned Transformer-Modelle und vortrainierte Large Language Models (LLMs) im Hinblick auf ihre Fähigkeit zum Question Answering zu vergleichen. Dabei werden die spezifischen Vorteile und Herausforderungen beider Ansätze untersucht. Die zentralen Fragestellungen lauten:

- Wie leistungsfähig sind Fine-Tuned Modelle im Vergleich zu LLMs bei der Beantwortung von Fragen?
- Welche Unterschiede gibt es hinsichtlich ihrer Anpassungsfähigkeit und Anwendbarkeit in realen Szenarien?
- Welche Modelle liefern präzisere Ergebnisse bei der Evaluierung anhand definierter Metriken?

2.2 Verwendete Modelle

Für den Vergleich werden folgende Modelle genutzt:

1. **Fine-Tuned Transformer-Modelle:**
    - **DistilBERT**: Ein kompaktes und effizientes Modell, das speziell für QA-Aufgaben angepasst wird [1].
    - **T5**: Ein flexibles Modell, das durch Fine-Tuning auf den SQuAD-Datensatz für QA optimiert wird [2].

2. **Vortrainierte LLMs:**

    - **LLAMA 2 7b Chat**: Ein leistungsstarkes Modell, das über das Hugging Face Interface [4] genutzt wird.
    - **ChatGPT**: Ein LLM, das über das Web-Interface von OpenAI verfügbar ist [3].

2.3 SQuAD-Datensatz

Der Stanford Question Answering Dataset (SQuAD) ist einer der meistgenutzten Datensätze für das Training und die Evaluierung von Question-Answering-Modellen.

- **Überblick:** SQuAD Version 1.0 umfasst mehr als 100.000 Fragen, die auf Wikipedia-Artikeln basieren. Jede Frage ist mit einem spezifischen Kontext verknüpft, aus dem die Antwort abgeleitet werden kann.
- **Aufbau:** Jede Frage hat einen zugeordneten Kontextabschnitt, aus dem die Antwort durch Extraktion eines oder mehrerer Wörter gefunden wird.

Zusätzlich wird in dieser Arbeit ein kleinerer, zufällig ausgewählter Test-Datensatz (TD-B) mit 100 Fragen erstellt, um die Leistung der Modelle in einem reduzierten Szenario zu bewerten.

2.4 Methodik

Die Untersuchung basiert auf einer systematischen Evaluierung der Modelle. Dazu gehören:

- Das Training und Testen der Fine-Tuned Modelle auf dem SQuAD-Datensatz.
- Die Erstellung eines kleineren Test-Datensatzes (TD-B) mit 100 Fragen, der für den Vergleich zwischen LLMs und Fine-Tuned Modellen verwendet wird.
- Die Evaluierung der Modelle anhand definierter Metriken wie F1 Score und Exact Match Score.
- Die Analyse von Eingaben und Ausgaben der LLMs, um irrelevante Inhalte zu entfernen und eine präzise Bewertung sicherzustellen.

2.5 Evaluierung

2.5.1 Methoden

Die Modelle werden mit folgenden Metriken evaluiert:

- **F1 Score**: Der F1-Score ist ein statistisches Maß, das zur Bewertung der Genauigkeit eines Modells verwendet wird, indem der harmonische Mittelwert von Präzision und Recall berechnet wird. Der F1-Score wird definiert durch
$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$
- **Exact Match Score**: ist eine strenge Bewertungsmetrik, die misst, ob die Antwort eines Modells genau mit einer der gegebenen Wahrheiten übereinstimmt. Er wird häufig verwendet, da er einen klaren, binären Indikator für die Genauigkeit eines Modells bei der Reproduktion präziser Antworten bietet, was die Bewertung und den Vergleich der Leistung verschiedener Modelle erleichtert.

2.5.2 Vorgehen

- Fine-Tuned Modelle werden auf TD-A und TD-B getestet.
- Vortrainierte LLMs werden auf TD-B getestet.

Alle Eingaben und Ausgaben werden gespeichert. Bei vortrainierten LLMs wird darauf geachtet, irrelevante Texte oder Einleitungen zu entfernen, um eine präzise Evaluierung sicherzustellen.

3 Training der Modelle

Aufgrund begrenzter Rechenressourcen wurde die Strategie gewählt, die besten Hyperparameter mithilfe von **Optuna** zu suchen. Dabei wurde der Fokus auf zwei wesentliche Parameter gelegt: *Batch-Größe* und *Lernrate*. Andere Hyperparameter konnten in dieser Arbeit nicht untersucht werden.

3.1 Hyperparameter-Optimierung mit Optuna

Optuna ist ein leistungsfähiges Framework für die automatisierte Hyperparameter-Optimierung. Für jedes Modell wurde eine separate *Optuna-Studie* durchgeführt, die darauf abzielte, den F1-Score auf dem Validierungsdatensatz zu maximieren. Dabei wurden die folgenden Einstellungen berücksichtigt:

- **Batch-Grösse:** Mögliche Werte: {16, 32, 64}.
- **Lernrate:** Suchraum: $[10^{-6}, 10^{-4}]$ im logarithmischen Maßstab.
- **Epochen:** Für T5 wurde die Anzahl der Epochen auf 20 festgelegt, während DistilBERT mit bis zu 20 Epochen trainiert wurde. Ein *Early Stopping* Mechanismus wurde verwendet, um das Training frühzeitig abzubrechen, falls keine Verbesserung der Validierungsmetriken festgestellt wurde.

Während jedes Optuna-Trials wurde das Modell auf dem Trainingsdatensatz trainiert und anschließend auf dem Validierungsdatensatz evaluiert. Die Modelle wurden anhand der Metriken *F1-Score* und *Exact Match Score* bewertet. Zusätzlich wurde für jeden Trial eine Tabelle der Trainings- und Validierungsverluste gespeichert. Dies erlaubte eine detaillierte Analyse, um sicherzustellen, dass die besten Parameter keine überanpassung (*Overfitting*) zeigten.

3.2 Durchgeführte Experimente

Für die Modelle **T5** und **DistilBERT** wurden die folgenden Schritte durchgeführt:

1. **T5:** Die besten Parameter wurden auf Basis einer Optuna-Studie mit 20 Trials ermittelt. Als Suchraum wurden Lernraten zwischen 10^{-6} und 10^{-4} sowie Batch-Größen von {16, 32} gewählt.
2. **DistilBERT:** Eine separate Optuna-Studie mit 20 Trials wurde durchgeführt, wobei die gleichen Parameter wie bei T5 untersucht wurden.

Die besten Hyperparameter wurden anschließend verwendet, um die Modelle final zu trainieren und ihre Leistung auf den Testdatensätzen zu evaluieren.

3.3 Einschränkungen und Entscheidungen

- **Begrenzte Ressourcen:** Aufgrund der begrenzten Rechenressourcen konnten nicht alle Hyperparameter untersucht werden. Daher wurde der Fokus auf die beiden wichtigsten Parameter gelegt: Batch-Größe und Lernrate.
- **Optimierungsziel:** Das Ziel der Optuna-Studien war die Maximierung des F1-Scores. Andere Metriken, wie z. B. Trainingszeit oder Speicherverbrauch, wurden nicht berücksichtigt.
- **Overfitting vermeiden:** Durch die Analyse der Trainings- und Validierungsverluste konnte sichergestellt werden, dass die gewählten Parameter keine überanpassung verursachten.

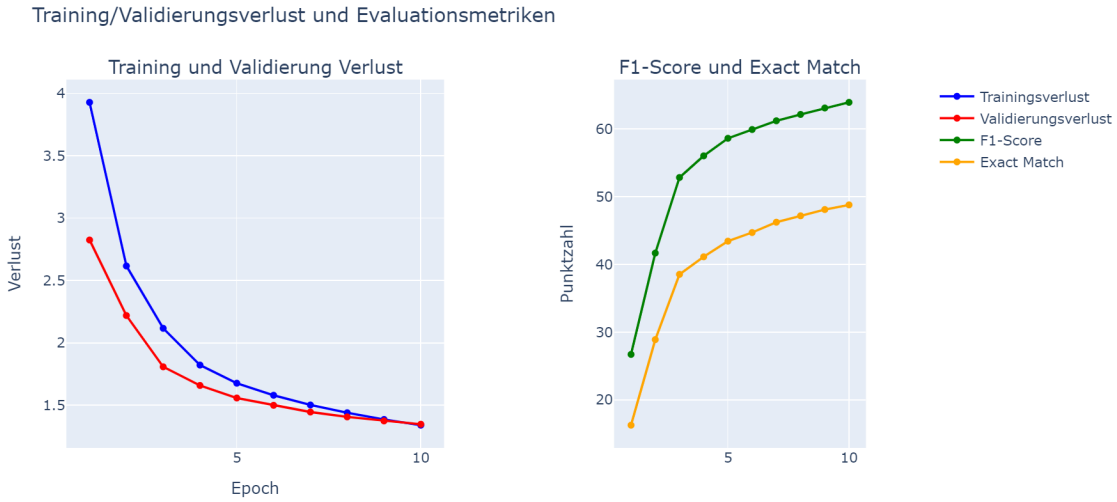
4 Ergebnisse des Modelltrainings

4.1 Auswahl der besten Optuna-Trials

Die Auswahl der finalen Hyperparameter für die Modelle wurde basierend auf den Optuna-Studien getroffen. Dabei wurden nicht nur die besten F1-Scores berücksichtigt, sondern auch die Trainings- und Validierungsverluste analysiert, um überanpassung (*Overfitting*) zu vermeiden. Die Details zu den ausgewählten Trials und den entsprechenden Modellen sind wie folgt:

4.1.1 DistilBERT

Für das Modell **DistilBERT** wurde der *Trial 18* der Optuna-Studie ausgewählt. Obwohl dieser Trial nicht den höchsten F1-Score aufwies, zeigte er im Vergleich zu den anderen Trials weniger überanpassung. Dies ist im folgenden Grafik ersichtlich:

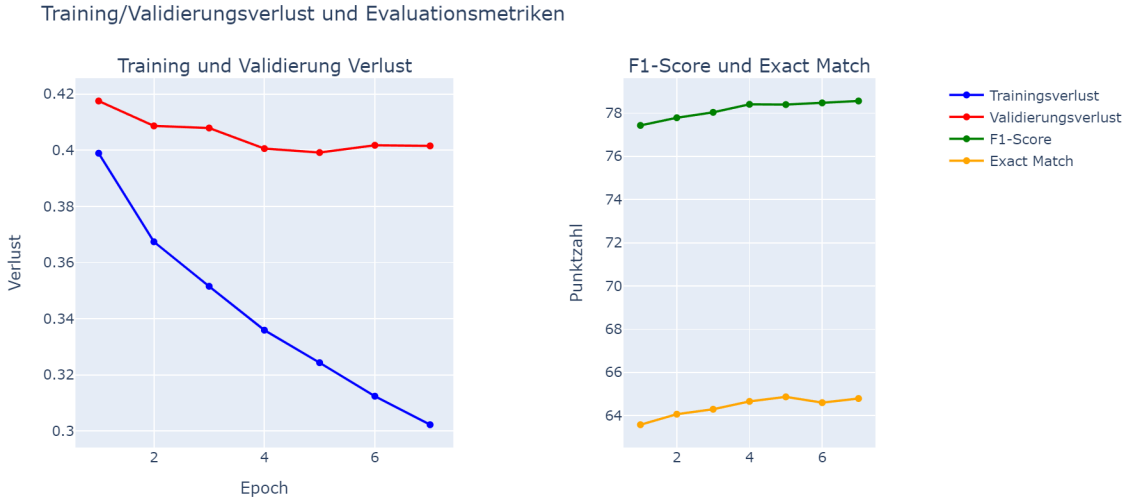


Die finalen Hyperparameter für das DistilBERT-Modell sind:

- **Lernrate:** 1.2784112998482085e-06
- **Batch-Größe:** 64

4.1.2 T5

Für das Modell **T5** wurde der *Trial 6* ausgewählt. Dieser Trial zeigte nicht nur einen hohen F1-Score, sondern auch stabile Trainings- und Validierungsverluste, die auf eine gute Generalisierungsfähigkeit hindeuten.



Die finalen Hyperparameter für das T5-Modell sind:

- **Lernrate:** 2.6762693430650424e-05
- **Batch-Größe:** 32

5 Analyse der Modellleistungen

5.1 Antwortgenerierung

Für die Modelle **ChatGPT-4o-min** und **Llama-2-7b** wurden folgende Verfahren zur Antwortgenerierung verwendet:

- **ChatGPT-4o-min:** Die API von OpenAI wurde genutzt, um Antworten basierend auf den gestellten Fragen und Kontexten zu generieren.

- **Llama-2-7b:** Das Modell wurde über die `pipeline`-Bibliothek von Hugging Face aufgerufen, um Antworten zu generieren.

Um sicherzustellen, dass die Modelle keine zusätzlichen Kontexte oder Erklärungen in ihre Antworten einfügen, wurde folgende Anweisung an die Fragen hinzugefügt:

> *"Provide only the exact answer to the questions from the SQuAD dataset, ensuring no additional context, elaboration, or supplementary details are included. Your responses must strictly adhere to the relevant passages from the dataset, offering nothing beyond the precise text necessary to answer the question. Under no circumstances should you incorporate external information or expand on details that are not present in the dataset."*

5.2 Ergebnisse der Modelle und Interpretation der Ergebnisse

Basierend auf den Ergebnissen der Modelle, insbesondere der vergleichsweise niedrigeren Leistung von ChatGPT-4o-min, erscheint eine detailliertere Analyse notwendig.

	Exact Match Score	F1 Score
Llama_2_7_b	61.0	79.759531
T5	64.0	78.169473
Distilbert	51.0	66.654936
chatgpt-4o-mini	61.0	79.759531

T5 erzielt mit 64% den höchsten Exact Match Score (EM), was auf seine Fähigkeit hinweist, exakte Antworten zu liefern. Llama_2_7_b und chatgpt-4o-mini folgen mit 61% und sind in der präzisen Beantwortung ähnlich. Distilbert schneidet mit

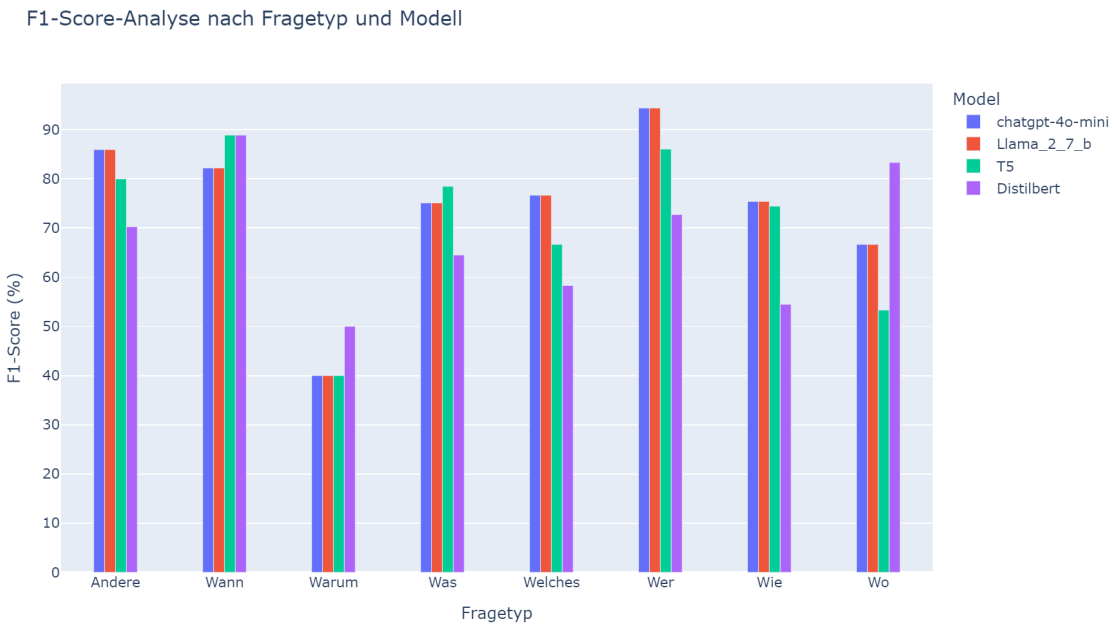
51% deutlich schlechter ab. Beim F1-Score führen Llama_2_7_b und chatgpt-4o-mini mit 79.76%, gefolgt von T5 mit 78.17%. Distilbert bleibt mit 66.65% auch hier hinten. T5 zeigt eine stärkere Präzision, während Llama_2_7_b und chatgpt-4o-mini ausgewogen gute Ergebnisse in beiden Metriken liefern. Llama_2_7_b und chatgpt-4o-mini eignen sich gut für generative Aufgaben, T5 für präzise Aufgaben, während Distilbert in beiden Bereichen schwächer abschneidet.

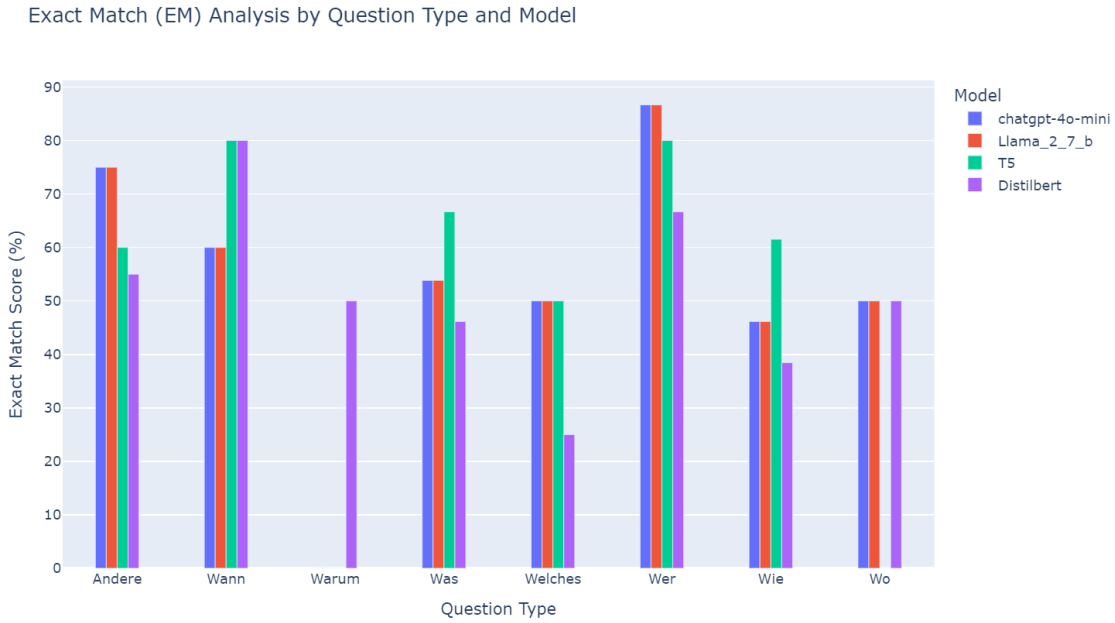
6 Vertiefte Analyse

Um einen umfassenderen Einblick in die Fähigkeiten der Modelle im Bereich Question Answering zu erhalten, wurden die EM-Scores auf Grundlage der Fragearten im SQuAD-Datensatz analysiert. Die Fragen wurden in die folgenden Kategorien eingeteilt:

- **Pronomen und Fragewörter:** *"Was", "Wie", "Wann", "Wo", "Wer", "Welches", "Warum", und Andere.*

Die Ergebnisse dieser detaillierten Analyse sind in den folgenden Grafiken dargestellt:





Exact Match (EM) Analyse: Die Analyse der EM-Werte nach Fragearten zeigt, dass die Modelle bei klar faktischen Fragen (z.B. "Wann", und "Wer") besser abschneiden, während bei erklärungsbasierten Fragen (z.B. "Warum") deutlich schwächere Leistungen erzielt werden.
F1-Score Analyse: Die F1-Analyse verdeutlicht, dass "Warum", und "Wie", Fragen, die oft komplexere und kontextuelle Antworten erfordern, eine Herausforderung für alle Modelle darstellen.

Die detaillierte Analyse zeigt:

Question_Type	Exact_Match	F1_Score
Andere	66.250000	80.542576
Wann	70.000000	85.555556
Warum	12.500000	42.500000
Was	55.128205	73.305098
Welches	43.750000	69.583333
Wer	80.000000	86.893939
Wie	48.076923	69.947403
Wo	37.500000	67.500000

- **Faktische Fragen:**
  - *"Wann", "Wer"*: Diese Fragearten erzielten die höchsten EM- und F1-Werte. Zum Beispiel erreichte "Wer"durchschnittlich einen F1-Score von 86.89%, wobei ChatGPT-4o-mini hier mit 94.39% besonders stark abschnitt.
  - *"Wo"*: Diese Fragen zeigten gemischte Ergebnisse mit durchschnittlichen F1-Werten von 67.50%, was darauf hindeutet, dass sie je nach Kontext schwieriger zu beantworten sind.
- **Komplexere Fragen:**
  - *"Warum"*: Dieser Fragetyp wies die niedrigsten Werte auf (F1: 42.50%, EM: 12.50%). DistilBERT erzielte hier einen F1-Score von 50%, konnte jedoch keine exakten übereinstimmungen liefern.

  - *"Wie"*: Diese Fragen zeigten ebenfalls eine Herausforderung, mit einem durchschnittlichen F1-Score von 69.95%. T5 schnitt hier mit einem F1-Score von 75.42% am besten ab.

- **Modellunterschiede:**
  - *ChatGPT-4o-mini*: Starke Leistung bei offenen und faktischen Fragen, insbesondere bei "Weründ Ändere"(F1: 85.95%).
  - *T5*: Zeigte hohe F1-Werte bei "WasFragen (78.47%) und gute Konsistenz bei komplexeren Fragen wie "Wie".
  - *DistilBERT*: Schwächer bei erklärungsbasierten Fragen, aber solide bei Ändere"(F1: 67.82%).
  - *Llama-2-7b*: Konsistente Leistung ähnlich zu ChatGPT-4o-mini, insbesondere bei faktischen Fragen.

7 Limitierungen

Trotz der umfassenden Analyse und der Verwendung moderner Modelle wie **DistilBERT**, **T5**, **ChatGPT-4o-min** und **Llama-2-7b** gibt es einige wesentliche Einschränkungen, die die Ergebnisse dieser Studienarbeit beeinflussen haben:

7.1 Rechenressourcen

Die verfügbaren Rechenressourcen waren begrenzt, was folgende Auswirkungen hatte:

- Nur eine beschränkte Anzahl an Hyperparametern (Batch-Größe und Lernrate) konnte mit **Optuna** optimiert werden.
- Die Anzahl der Optuna-Trials wurde auf 20 begrenzt, was die Exploration des Hyperparameter-Raums einschränkte.

7.2 Datensatzumfang

Obwohl der SQuAD-Datensatz eine bewährte Grundlage für Question Answering-Modelle bietet, gibt es auch hier Einschränkungen:

- Es wurde ein kleinerer, zufällig ausgewählter Testdatensatz (TD-B) mit 100 Fragen erstellt, um die Leistung der Modelle zu vergleichen. Ein größerer Datensatz hätte möglicherweise differenziertere Ergebnisse geliefert.
- Die Auswertung basiert ausschließlich auf SQuAD und berücksichtigt keine anderen Domänen oder Anwendungsfälle.

7.3 Modelleinschränkungen

Die gewählten Modelle haben eigene Grenzen:

- Die Fine-Tuned-Modelle (DistilBERT und T5) sind stark auf den Trainingsdatensatz angewiesen und könnten in Domänen mit abweichendem Kontext schlechter abschneiden.
- Die vortrainierten LLMs (ChatGPT und Llama-2-7b) haben generative Fähigkeiten, die manchmal irrelevante oder nicht präzise Antworten liefern.

8 Fazit und Zusammenfassung

In dieser Studienarbeit wurden Fine-Tuned Transformer-Modelle (**DistilBERT**, **T5**) und vortrainierte Large Language Models (**ChatGPT-4o-min**, **Llama-2-7b**) im Hinblick auf ihre Fähigkeiten im Bereich *Question Answering (QA)* untersucht. Die Ergebnisse zeigen:

- **Fine-Tuned Modelle:** Sie haben bei QA-Aufgaben auf dem spezifischen SQuAD-Datensatz eine starke Leistung gezeigt. Die Anpassung an den Trainingsdatensatz macht sie jedoch weniger flexibel in anderen Kontexten.
- **Vortrainierte LLMs:** Diese Modelle zeigten eine robuste Leistung, insbesondere bei kontextbezogenen und offenen Fragen. Ihre generativen Fähigkeiten bergen jedoch das Risiko, irrelevante oder unpräzise Informationen zu liefern.

Die detaillierte Analyse der Fragearten verdeutlichte zusätzliche Aspekte:

- Faktische Fragen wie "Wann", und "Wer", sind einfacher zu beantworten und erzielten höhere Scores.
- Erklärungsbasierte Fragen wie "Warum", stellen alle Modelle vor Herausforderungen.
- T5 zeigte starke Leistungen bei "Was", Fragen, während ChatGPT-4o-mini bei offenen Fragen konsistenter war.

Fine-Tuned-Modelle wie DistilBERT und T5 zeigen bei spezifischen QA-Aufgaben, insbesondere auf Datensätzen wie SQuAD, eine hohe Leistung und Präzision, da sie gezielt auf diese Aufgaben trainiert wurden. Im Vergleich dazu bieten vortrainierte LLMs wie ChatGPT-4o-mini und Llama-2-7b eine größere Anpassungsfähigkeit und Robustheit, insbesondere bei offenen und kontextabhängigen Fragen, jedoch oft auf Kosten der Präzision bei faktenbasierten Fragen. Fine-Tuned-Modelle sind

daher besonders für kontrollierte Umgebungen und spezialisierte Anwendungsfälle geeignet, während LLMs vielseitiger in realen Szenarien eingesetzt werden können, wo die Vielfalt der Fragestellungen und Kontexte größer ist. Bei der Evaluierung anhand definierter Metriken wie Exact Match (EM) und F1-Score liefern Fine-Tuned-Modelle präzisere Ergebnisse für klare, faktenbasierte Fragen, während LLMs bei erklärungsbasierten oder komplexeren Fragen eine höhere Flexibilität zeigen.

9 Quellenverzeichnis

- [1] huggingface DistilBertForQuestionAnswering
- [2] huggingface T5ForQuestionAnswering.
- [3] openai Chatgpt
- [4] Hugging Face Llama-2-7b-chat-hf
- [5] arxiv
- [6] medium guide to question answering with t5 and pytorch