



Hochschule für Technik
und Wirtschaft Berlin

University of Applied Sciences

Evaluierung von Retrieval-Augmented Generation und Fine-Tuning in der Wissensrepräsentation: Ein Vergleich mit Großen Sprachmodellen

Bachelorarbeit

Name des Studiengangs

Angewandte Informatik

Fachbereich 4

vorgelegt von

Florian Symmank

Datum:

Berlin, 09.08.2024

Erstgutachter: Prof. Dr. Ermakova

Zweitgutachter: Carsten Tittel

Abstrakt

Diese Bachelorarbeit befasst sich mit der Evaluation von Retrieval-Augmented Generation (RAG) und Fine-Tuning (FT) in der Wissensrepräsentation und vergleicht diese Methoden durch den Einsatz von Large Language Modellen (LLMs). Während konventionelle Suchmethoden und standardisierte Repräsentationen in Wissensdatenbanken bisher vorherrschten, bieten die Fortschritte im Bereich der großen Sprachmodelle (LLMs) neue Möglichkeiten der Interaktion auf einer höheren Sprachebene. Diese Arbeit untersucht die Eignung von RAG und FT im Vergleich zu nicht angepassten LLMs bei der Wissensrepräsentation auf einem typischen Wiki-Datensatz. Hierbei werden grundlegende Technologien und Verfahren näher erläutert, bestehende Studien analysiert und experimentelle Designs skizziert. Die Evaluation erfolgt durch die Messung von BLEU- und ROUGE-Scores. Es zeigt sich, dass RAG in Kombination mit Basismodellen besonders effektiv ist, während Fine-Tuning allein weniger Erfolg zeigt als erwartet. Der Einsatz von größeren Modellen führt zu besseren Ergebnissen. Die Untersuchung liefert wertvolle Erkenntnisse für die zukünftige Entwicklung und Anwendung von LLMs in der Wissensrepräsentation, insbesondere in dynamischen Umgebungen, in denen aktuelle und genaue Daten entscheidend sind. Abschließend werden Einschränkungen und Möglichkeiten für zukünftige Arbeiten diskutiert, um ein umfassenderes Bild der Leistungsfähigkeit und Effizienz dieser Methoden zu erhalten.

Inhaltsverzeichnis

Einführung	1
Problemstellung	1
Zielsetzung/Forschungsfragen	1
Aufbau der Arbeit	1
Grundlagen	2
Fine-Tuning	3
Retrieval-Augmented Generation	4
Related Work	5
Design von Experimenten	5
Modellauswahl	6
Datensatzbeschreibung	7
Werkzeug/Software	7
Trainingsmethoden	8
Anwendung Retrieval Augmented Generation	8
Evaluationsmetriken	9
Auswertung	11
Durchgeführte Evaluationen	11
Übersicht	12
Diskussion	21
Einschränkungen	21
Zukünftige Arbeiten	22
Zusammenfassung	22
Quellenverzeichnis	23
Abkürzungsverzeichnis	26
Anhang	27
Eidesstattliche Versicherung	44

Einführung

Unternehmen verfügen häufig über umfangreiche Wissensdatenbanken, die bislang hauptsächlich durch konventionelle Suchmethoden und standardisierte Repräsentationen zugänglich gemacht werden. Allerdings ergeben sich durch den exponentiellen Fortschritt und die zunehmende Verbreitung großer Sprachmodelle (Large Language Models, LLMs) nunmehr neue Möglichkeiten. Es wird immer interessanter, eigens trainierte und speziell an spezifische Aufgaben angepasste LLMs zu nutzen, um direkter und auf einer höheren Sprachebene mit diesen Wissensdatenbanken zu interagieren [1]. Neben der kontinuierlichen Weiterentwicklung der LLMs hat sich auch die Hardware, die für den Betrieb dieser Modelle notwendig ist, erheblich weiterentwickelt und ist inzwischen wesentlich zugänglicher geworden [2]. Der Einsatz von LLMs ist nicht ohne Überwindung von Herausforderungen möglich. Zu den bekannten Herausforderungen von LLMs gehören unter anderem die Missrepräsentation von Informationen [3] oder die Wiedergabe veralteter Informationen. Um diesen Problemen entgegenzuwirken, können LLMs mittels Fine-Tuning auf Domänenspezifisches Wissen angepasst werden. Änderungen am Domänenwissen ziehen allerdings ein kosten- und zeitaufwendiges Neu-Trainieren des LLMs nach sich [4], [5], falls das LLM darauf zugreifen soll. Eine weitere Möglichkeit LLMs Zugriff auf die neuesten und relevantesten Daten zu bieten, ist Retrieval-Augmented Generation (RAG) [6]. Hierbei werden kontext-relevante Dokumente und Informationen direkt bei der Nutzereingabe übergeben, ohne dass das LLM neu trainiert werden muss.

Problemstellung

In dieser Arbeit soll geklärt werden, wie geeignet ein LLM mit RAG gegenüber einem Fine-Tuned LLM und einem nicht angepassten LLM bei der Wissensrepräsentation auf einem typischen Wiki-Datensatz, wie er in verschiedenen Organisationen und Unternehmen zu finden, ist.

Zielsetzung/Forschungsfragen

- Ist ein LLM mit RAG oder ein fine-tuned LLM besser geeignet, um Informationen aus einem typischen Wiki-Datensatz effektiv zu repräsentieren?
- Was sind die Vor- und Nachteile der jeweiligen Trainingsmethode?
- Welchen Mehrwert bietet die Kombination von RAG und Fine-Tuning für die Wissensrepräsentation in LLMs im Vergleich zur Verwendung jeder Methode einzeln?
- Welche LLMs eignen sich dafür besonders?
- Welchen Einfluss hat die Parameteranzahl?

Aufbau der Arbeit

Im ersten Teil der Arbeit werden grundlegende Technologien und Verfahren näher erläutert und definiert. Es wird unter anderem erläutert, was ein Large Language Model ist, welches die Einsatzzwecke sind und wie es sich von anderen KI-Systemen abgrenzt. Es wird näher auf die beiden Methoden Fine-Tuning und RAG zum Erweitern des Domänenwissens eingegangen. Welche Vor- und Nachteile bestehen, wie wird das Wissen erweitert?

In der Sektion "Related Work" wird auf die Analyse bestehender Studien eingegangen, die zu unterschiedlichen Schlussfolgerungen hinsichtlich der Wirksamkeit verschiedener Techniken in LLMs kommen.

Im Teil Design von Experimenten wird auf die Methodik des Experiments und Trainings, die verwendeten Daten, sowie auf die Evaluationsmetriken, ROUGE und BLEU, und andere zugehörige Software eingegangen.

Neben der Auswertung der Versuche wird auf Einschränkungen der Arbeit und den Ausblick auf zukünftige Arbeiten eingegangen.

Grundlagen

Ein Large Language Model (LLM), auf Deutsch auch Sprachmodell genannt, ist ein maschinelles Lernmodell, das auf die Verarbeitung und Erzeugung natürlicher Sprache spezialisiert ist. Diese Modelle verwenden fortschrittliche Algorithmen, insbesondere neuronale Netzwerke, um Muster in umfangreichen Textdaten zu erkennen. Indem sie die Wahrscheinlichkeit des Auftretens einer Wortfolge in einem gegebenen Kontext vorhersagen, können LLMs menschenähnliche Texte generieren und die Wahrscheinlichkeit bestimmter Wortfolgen bewerten.

Das typische Aufgabenfeld der LLMs ist die Verarbeitung natürlicher Sprache (NLP), welches sich in zwei Kategorien einteilen lässt: Natural Language Understanding (NLU) und Natural Language Generation (NLG).

Transformer

Diese Arbeit beschäftigt sich mit Transformer Modellen, eine Architektur zum maschinellen Lernen, eingeführt in einem der grundlegenden Meilensteine für moderne multimodale generative KI und Natural Language Processing (NLP) "Attention is All you Need" [7]. Infolge der Einführung von Transformermodellen wurden Modell Architekturen wie Recurring Neural Network (RNN) und Convolutional Neural Networks (CNN) größtenteils aus dem NLP Umfeld verdrängt [8] und lassen sich auch nicht in aktuellen Benchmarks wie z.B. im HuggingFace LLM Leaderboard wiederfinden [9].

Neben Natural Language Processing lassen sich Transformer auch in anderen Gebieten einsetzen.

Natural Language Processing (NLP)		Computer Vision	Audioanwendungen	Multimodale Anwendungen
Natural Language Understanding (NLU)	Natural Language Generation (NLG)			
<ul style="list-style-type: none"> - Sentiment Analyse - Text Klassifikation - Natural language inference (NLI) 	<ul style="list-style-type: none"> - Zusammenfassung - Dialogue - Frage/Antwort - Übersetzung 	<ul style="list-style-type: none"> - Bildklassifizierung - Objekterkennung - Bilderzeugung - Videoverarbeitung 	<ul style="list-style-type: none"> - Spracherkennung - Sprachsynthese - Verbesserung der Sprache - Erzeugung von Musik 	<ul style="list-style-type: none"> - Visuelle Fragenbeantwortung - Visuelles Schlussfolgern - Erzeugung von Untertiteln - Übersetzung von Sprache in Text - Text-zu-Bild-Generierung

Tabelle 1: Anwendungen von Transformern [\[10\]](#)

Wie in der Tabelle "Anwendungen von Transformern" zu erkennen ist, können Transformer nicht nur Texte, sondern auch Bilder, Tonsignale und andere Datentypen verarbeiten. Diese Arbeit und die nachfolgenden Erläuterungen beschränken sich auf die NLP Möglichkeiten von Transformern.

Funktionsweise

Der Tokenisierungs- und Eingabeprozess bei Transformer ist von grundlegender Bedeutung für ihr Funktionsprinzip. Die Eingabedaten werden zuerst in eine Sequenz von Token umgewandelt, wobei jeder Token durch einen Vektor repräsentiert wird. Anschließend werden die Token in den Transformer eingespeist. Dieser Prozess der Tokenisierung ermöglicht es dem Transformer, eine Vielzahl von Datentypen zu verarbeiten. Im Gegensatz zu früheren Architekturen, die spezifisch für bestimmte Datentypen entwickelt wurden (z.B. CNNs für Bilder, RNNs für Sequenzen), erlauben es Transformatoren, verschiedene Datentypen in Form von Tokens einfach zu verarbeiten, ohne dass maßgeschneiderte Architekturen für jede Modalität benötigt werden. Zusätzlich ermöglicht die Tokenisierung, dass verschiedene Datentypen problemlos miteinander gemischt werden können, da sie in einen großen Satz von Tokens geworfen werden können.

Der Transformator besteht aus mehreren Schichten, wobei jede Schicht zwei wichtige Verarbeitungsstufen enthält. Die erste Stufe wird als Selbstaufmerksamkeit (engl. self-attention) bezeichnet, bei der der Transformator die Beziehungen zwischen den verschiedenen Token in der Sequenz betrachtet. Dies geschieht durch die Berechnung einer Aufmerksamkeitsmatrix, die erfasst, wie stark jeder Token in der Sequenz mit den anderen verwandt ist. Dadurch kann die Darstellung jedes Tokens basierend auf den anderen Token um ihn herum verfeinert werden. Die zweite Phase in jeder Transformer-Schicht ist ein mehrschichtiges Perzeptron, ein weiteres künstliches neuronales Netzwerk, (MLP), das die Merkmale jedes Tokens einzeln bearbeitet und die Darstellung weiter verfeinert [11].

Fine-Tuning

In der Pre-Trainingsphase lernt das Modell, dann oft als Foundation Model (FM) oder Pre-Trained Model (PTM) bezeichnet, zunächst aus einem großen Korpus unmarkierter Daten. Fine-Tuning bezieht sich auf die Verfeinerung eines vortrainierten Modells mit Hilfe von beschrifteten Daten, die für eine nachgelagerte Aufgabe spezifisch sind, um seine Leistung für diese spezielle Anwendung zu verbessern. Aus einem einzigen FM können so mehrere fein abgestimmte Modelle abgeleitet werden. Fine-Tuning wird auch häufig verwendet, um verändertes Domänenwissen zu berücksichtigen. Ein weiterer Grund zur Aufteilung der Trainingsphasen besteht darin, dass es schlicht kosteneffizienter ist. Pre-Trainings in großem Maßstab erfordert große Mengen an Berechnungen, die sehr energieintensiv sind [12], und können in den meisten Fällen nur durch gut finanzierte Industrieunternehmen durchgeführt werden [13].

Selbst das Fine-Tuning solcher FMs erfordert erheblichen Rechenaufwand. Modelle wie GPT-3 mit 175 Milliarden Parametern, GPT-4 soll weitaus mehr besitzen, stellen alleine aufgrund der schieren Zahl an trainierbaren Parametern eine Herausforderung des Fine-Tunings dar. Die verbundenen Kosten aufgrund des benötigten Speicherbedarfs, der GPU-Rechenleistung und Energieverbrauchs sind enorm. Um diese Problematik zu bewältigen, wurden verschiedene Verfahren entwickelt, die die Anzahl der zu trainierenden Parameter reduzieren. LoRA, oder low-rank adaptation, führt eine neue trainierbare Schicht in Transformer ein, welche, im Falle von GPT-3, die Anzahl der Parameter um das 10.000-fache und den GPU-Bedarf um das Dreifache reduziert. Neben der Vereinfachung des Fine-Tunings entsteht durch LoRA keine weitere Inferenzlatenz, wie es bei anderen Adaptionen der Fall ist. [14] Eine Weiterentwicklung von LoRA mit dem Ziel die Speicheranforderungen beim Fine-Tuning noch weiter zu reduzieren, ist LoRA für Quantized LLMs (QLoRA). Mit QLoRA ist es möglich, LLMs mit 65 Milliarden Parametern auf einer einzigen GPU fein abzustimmen. Der Speicherbedarf sinkt ohne

Verschlechterung der Laufzeit oder Vorhersageleistung im Vergleich zu einem 16-Bit fein abgestimmten Modell von über 780 GB auf unter 48 GB [15].

Retrieval-Augmented Generation

Retrieval-Augmented Generation (RAG) ist ein innovativer Ansatz, der darauf abzielt, die Qualität und Genauigkeit der durch Large Language Models generierten Texte durch die Integration von externen Information-Retrieval-Mechanismen erheblich zu verbessern. Während herkömmliche LLMs zwar äußerst leistungsfähig sind, bleiben sie durch die Daten, auf denen sie trainiert wurden, limitiert und können daher manchmal veraltete oder ungenaue Informationen liefern. Diese Einschränkungen mindern den Nutzen solcher Modelle in dynamischen Umgebungen, in denen aktuelle und genaue Daten entscheidend sind. RAG setzt hier an und kombiniert die Generierungsfähigkeiten von LLMs mit der Echtzeitfähigkeit, relevante Daten aus externen Ressourcen abzurufen. Dieser Ansatz lässt sich in drei primäre Schritte unterteilen:

- Document Retrieval
- Context Enrichment
- Training des Retrievers und Generators

Durch diese mehrstufige Vorgehensweise wird sichergestellt, dass die generierten Texte sowohl inhaltlich korrekt als auch aktuell sind und somit den Anforderungen der Nutzer besser entsprechen.

In der Document-Retrieval-Phase werden passende Dokumente zum gegebenen Input gesucht. Historisch wurden hierbei "Sparse Vector Model"-Verfahren wie BM25 [16, S. 25] und TF-IDF [17] verwendet, um relevante Dokumente oder Dokumentpassagen zu identifizieren. Diese Methoden nutzen nur wenige Wörter oder Token für die Vektorrepräsentation der Dokumente. Aktuelle Entwicklungen ermöglichen jedoch die Verwendung von "Dense Encodings" wie Dense Passage Retrieval (DPR) [18]. Diese Ansätze, die auf Transformern basieren, können den Kontext der Textpassage besser aufnehmen und repräsentieren. Die entstehende Vektorrepräsentation besteht aus vielen nicht-null Fließkommazahlen, was zu einer präziseren Abbildungen führt. Durch den Vergleich der gefundenen Dokumente mit dem Input — meist eine Query (zu Deutsch: Anfrage) an das Large Language Model — können so die relevantesten Dokumente identifiziert werden.

Die für die Anfrage als relevant eingestuften Dokumente werden in die Anfrage an das LLM übernommen und dienen somit als Kontext. In "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks" [19] werden zwei Methoden zur Handhabung der Dokumente und der Suche nach diesen vorgestellt. Die einfache Methode, "RAG-Sequence", sucht die Dokumente lediglich zu Beginn der Anfrage. Im Gegensatz dazu sucht "RAG-Token" nach n erzeugten Token erneut nach Dokumenten, um weitere, nun relevant gewordene Dokumente in die Textgenerierung zu integrieren. Ein weiterer Ansatz, bekannt als RETRO [20], integriert die Dokumente direkt in die Sub-Layer des Transformers.

Einige dieser Ansätze beinhalten das Fine-Tuning der einzelnen Komponenten, wie des Retrievers und Generators, oder gar des gesamten Systems im End-to-End-Verfahren [19], [21]. Andere Verfahren, wie beispielsweise das oben erwähnte RETRO, verzichten explizit auf das Fine-Tuning, was als Frozen RAG bezeichnet wird.

Related Work

Die Auswertung ähnlicher Arbeiten hat kein eindeutiges Bild ergeben. Eine Studie, die verschiedene Dialogtypen – Open-Domain, Knowledge-Grounded, Task-Oriented und Question Answering – untersucht hat, kommt zu dem Schluss, dass es keine universell beste Technik gibt und je nach Large Language Model sowie Dialogtyp Unterschiede bestehen [22].

Eine andere Studie, die Retrieval-Augmented Generation und Fine-Tuning im landwirtschaftlichen Bereich für LLMs untersucht, gelangt zu dem Schluss, dass RAG und FT unterschiedliche Vorteile bieten. RAG kann durch kontextrelevante Zusatzinformationen sehr effektiv werden, während FT präzisere und kürzere Antworten liefert [23].

Im Gegensatz dazu kommt „Fine-Tuning or Retrieval? Comparing Knowledge Injection in LLMs“, welche die Felder Anatomie, Astronomie, College-Biologie, Chemie und Prähistorie aus dem Massively Multilingual Language Understanding Evaluation (MMLU) Benchmark untersucht, zu dem Ergebnis, dass RAG ohne FT bessere Ergebnisse liefert als RAG mit FT [24].

Eine weitere Studie, die FT und RAG für weniger bekanntes Wissen untersucht, kommt hingegen zu dem Schluss, dass FT gegenüber RAG unterlegen ist. RAG wiederum wird von einer Kombination aus FT und RAG übertroffen [25].

Design von Experimenten

Im nächsten Kapitel werden die einzelnen Komponenten des Versuchsaufbaus detailliert erläutert. Abbildung 1 zeigt den Ablauf des Experiments unter Verwendung eines fein abgestimmten LLM und RAG.

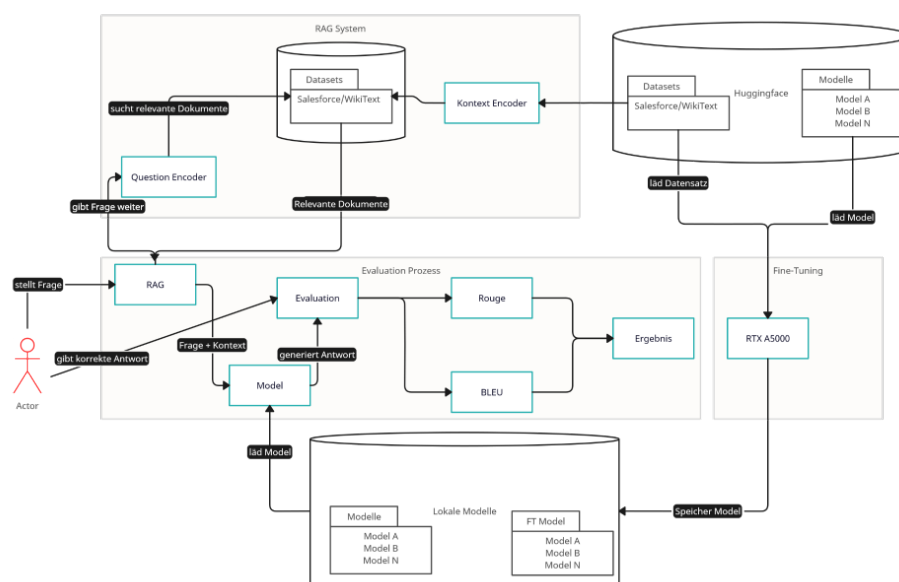


Abbildung 1: Prozessbild mit RAG und FT

Modellauswahl

Alle nachfolgenden Modelle wurden aus dem Huggingface Open LLM Leaderboard [9] ausgewählt. Diese Modelle schnitten im Vergleich zu ähnlichen Modellen der gleichen Größenordnung laut Leaderboard gut ab oder es gibt ein verwandtes Modell desselben Autors, das gute Ergebnisse erzielte. Dies ermöglicht, Vergleiche innerhalb der Modellfamilie zu ziehen. Jedes Modell ist für einfache Interaktionen mithilfe von "Instruct" oder "Chat" Fine-Tuning angepasst, um einen fairen Vergleich zu gewährleisten. Zudem sind die Modelle nicht quantisiert. Modelle, die größer als 8B sind, können aufgrund von Hardware-Limitierungen nicht ausgeführt werden. Modelle größer als 2B konnten nicht trainiert werden.

0B - 2B Parameter Modelle

Alle Modelle in dieser Kategorie besitzen maximal 2 Milliarden Parameter.

Wie alle Qwen2 Sprachmodelle basiert das Qwen/Qwen2-0.5B-Instruct Modell auf der Transformer-Architektur und verwendet die SwiGLU-Aktivierung, den Attention QKV Bias und die Group Query Attention. Mit einer Modellgröße von 494 Millionen Parametern verwendet es den Tensor Typ BF16 (Brainfloat-16) [26]. Ein weiteres Mitglied der Qwen2-Familie ist das Qwen/Qwen2-1.5B-Instruct Modell, das 1,54 Milliarden Parameter aufweist und ebenfalls den BF16-Tensor Typ nutzt [27].

Analog zur Qwen2-Serie basiert auch die Qwen1.5-Familie auf der Transformer-Architektur und verwendet die SwiGLU-Aktivierung, den Attention QKV Bias sowie die Group Query Attention. Ergänzend dazu setzt diese Modellserie auf eine Kombination aus Sliding Window Attention und vollständiger Aufmerksamkeit. Das Qwen/Qwen1.5-0.5B-Chat Modell besitzt 620 Millionen Parameter und nutzt den BF16-Tensor Typ [28], während das Qwen/Qwen1.5-1.8B-Chat Modell über 1,84 Milliarden Parameter verfügt und ebenfalls den BF16-Tensor Typ verwendet [29].

Das Modell internlm/internlm2-chat-1_8b hat 1,89 Milliarden Parameter und verwendet ebenfalls den BF16-Tensor Typ [30].

Stabilityai/stablelm-2-1_6b-chat ähnelt der LLaMA-Architektur [31] und ist ein Decoder-Sprachmodell mit 1,64 Milliarden Parametern, das den TF32-Tensor Typ nutzt [32].

HuggingFaceTB/SmolLM-1.7B-Instruct wurde auf dem Cosmo-Corpus, einem sorgfältig kuratierten und hochwertigen Trainingsdatensatz, trainiert. Dieses Modell besitzt 1,71 Milliarden Parameter und nutzt den BF16-Tensor Typ [33].

4B Parameter Modelle

Die Microsoft Phi-3-Reihe wurde auf dem Phi-3-Datensatz trainiert. Das Modell microsoft/Phi-3-mini-128k-instruct passt trotz seiner 3,82 Milliarden Parameter im BF16-Format nach 4-Bit Quantisierung problemlos auf aktuelle Smartphones und erreicht dabei ähnliche Leistungen wie das Mixtral 8x7B Modell [34], [35].

Qwen/Qwen1.5-4B-Chat ist ein Modell der Qwen1.5 Familie und verfügt über 3,95 Milliarden Parameter des Tensor Typs BF16 [36].

7B - 8B Parameter Modelle

Das Modell mistralai/Mistral-7B-Instruct-v0.3, welches über 7,25 Milliarden Parameter (BF16) verfügt, hat keinerlei Moderationsmechanismen und ist daher ein interessanter Kandidat [37].

Darüber hinaus bildet die Qwen-Sprachmodell Familie die Grundlage für die darauffolgenden Versionen Qwen1.5 und Qwen2. Besonders erwähnenswert ist das Modell Qwen/Qwen-7B-Chat, welches 7,72 Milliarden Parameter (BF16) besitzt [38].

Im weiteren Verlauf unserer Analyse betrachten wir auch das Modell Qwen/Qwen2-7B-Instruct, welches mit seinen 7,62 Milliarden Parametern (BF16) zur Qwen2-Sprachmodellfamilie gehört [39].

Im Zusammenhang mit den autoregressiven Sprachmodellen ist Llama 3 hervorzuheben. Dieses Modell verwendet eine optimierte Transformer-Architektur und kombiniert supervised fine-tuning (SFT) mit reinforcement learning durch menschliches Feedback (RLHF). Dadurch passt es sich besser an menschliche Präferenzen hinsichtlich Nützlichkeit und Sicherheit an. Das Modell meta-llama/Meta-Llama-3-8B-Instruct mit 8,03 Milliarden Parametern (BF16) ist die erste Version dieser Familie [40]. Interessanterweise gibt es ein Update mit dem Namen meta-llama/Meta-Llama-3.1-8B-Instruct, welches ebenfalls 8,03 Milliarden Parameter (BF16) umfasst, jedoch durch die Unterstützung mehrsprachiger Ein- und Ausgaben hervorsteicht [41]. Ein Vergleich dieser beiden Modelle kann spannende Einblicke in die Entwicklung und Weiterentwicklung der LLaMA 3 Modelle bieten.

Datensatzbeschreibung

Der Datensatz Salesforce/WikiText, genauer gesagt der Datensatz „wikitext-2-raw-1“ [42] wird sowohl während des Fine-Tuning als auch bei RAG und der anschließenden Evaluierung der generierten Texte verwendet.

Der WikiText-Datensatz wurde speziell entwickelt, um die Schwächen des früher weit verbreiteten Penn Treebank (PTB) Datensatzes zu überwinden. Der PTB-Datensatz hat ein kleines Vokabular, ist kleingeschrieben, hat keine Interpunktion und repräsentiert somit nicht den realen Sprachgebrauch. WikiText behebt diese Einschränkungen, indem er ein größeres Korpus mit originaler Groß- und Kleinschreibung, Interpunktion und einem größeren Vokabular bereitstellt, was es für Sprachmodellierungsaufgaben deutlich realistischer macht [43]. Seit seiner Einführung wird dieser Datensatz häufig zum Trainieren von umfangreichen Sprachmodellen (LLMs) verwendet. Wikitext-2-raw-v1 enthält die drei üblichen Datensplits: Training, Validierung und Test. Es besteht aus einem einzigen Merkmal, nämlich Text, und umfasst über 44.000 Datensätze. Ein Großteil der Texteinträge, etwa 67%, besitzt eine maximale Länge von 387 Zeichen. Weniger als 5% der Texte sind länger als 1400 Zeichen. Interessanterweise ist die Verteilung der Textlänge in den verschiedenen Datensplits nicht gleichmäßig. Alle Texte sind in englischer Sprache. Anhang 5 zeigt einen beispielhaften Textauszug.

Werkzeug/Software

In dieser Arbeit wurde ausschließlich mit Python 3.10 innerhalb einer JupyterHub-Umgebung gearbeitet. Es ist wichtig zu erwähnen, dass einige der verwendeten Python-Bibliotheken aus Performance-Gründen in C geschrieben sein können.

Alle Modelle wurden auf einer NVIDIA RTX A5000 mit 24,564 MiB VRAM trainiert. Um das Training zu vereinfachen und zu beschleunigen, kamen die Bibliotheken wie torch, flash-attn, accelerate und deepspeed zum Einsatz. Allerdings wurden die beiden letztgenannten Bibliotheken – accelerate und deepspeed – in dieser Arbeit nicht genutzt.

Besonders hervorzuheben ist, dass der Umgang mit LLMs durch die verschiedenen Huggingface-Transformer-Bibliotheken (🤗 Transformers) stark vereinfacht wurde.

Trainingsmethoden

Alle trainierten Modelle wurden einmal auf dem Salesforce/WikiText - Datensatz "wikitext-2-raw-1" [42] trainiert. Die verwendeten Parameter und Konfigurationen sind in Anhang 1 und Anhang 2 aufgeführt. Die Parameter entsprechen größtenteils den Standardeinstellungen, und das Fine-Tuning mit LoRA wurde über drei Epochen durchgeführt, gefolgt von Evaluationen nach jeder Epoche. Eine Änderung zu den Voreinstellungen ist, dass nach Abschluss des Trainings das beste Modell anhand der eval_loss-Funktion ausgewählt wurde. Zusätzlich wurde eine Warm-Up-Phase von 500 Schritten eingefügt, in der die Lernrate schrittweise auf den Startwert von $5e-5$ (0.00005) erhöht wurde, um anfängliche Optimierungsprobleme zu überwinden [44]. Das FT dauerte jeweils 90 bis 240 Minuten.

Das Finden der optimalen Fine-Tuning-Einstellungen ist nicht Teil dieser Arbeit, da viele Parameter beeinflussbar sind, welche nicht alle evaluiert werden konnten.

Anwendung Retrieval Augmented Generation

In der Anwendung von RAG wurden in der Document-Retrieval-Phase die Encoder-Modelle facebook/dpr-question_encoder-single-nq-base [44] und facebook/dpr-ctx_encoder-single-nq-base [45] verwendet. Beide Modelle nutzen Dense Passage Retrieval (DPR), um Texte zu vektorisieren und somit besser durchsuchen zu können.

Um den Eingabekontext zu erweitern, wird der Trainings-Split aus dem WikiText-Datensatz verwendet. Dies ist wichtig, da im Kontext einer Suche und Beantwortung in einer tatsächlichen Wissensdatenbank dieselben Daten erlernt, genutzt und optimalerweise wiedergegeben werden sollen, ähnlich wie beim FT Verfahren. Es wird das RAG-Sequence-Verfahren angewendet, bei dem relevante Dokumente nur einmal vor Beginn der Textgeneration gesucht werden.

Zunächst wird der Datensatz mithilfe des dpr-ctx_encoder-single-nq-base in die entsprechende Vektorrepräsentation der einzelnen Datenpunkte überführt. Zur Anreicherung des Kontexts wird die Benutzereingabe durch den dpr-question_encoder-single-nq-base kodiert. Anschließend können durch die Berechnung der Kosinusähnlichkeit zweier Vektoren die n (in dieser Arbeit immer $n=3$) relevantesten Dokumente gefunden werden. Diese Dokumente werden dann der Nutzerfrage als zusätzlicher Kontext angehängt. Um eine feinere Kontrolle über die als relevant eingestufteten Dokumente zu erhalten, kann ein Schwellenwert hinzugefügt werden, den die Kosinusähnlichkeit überschreiten muss, damit ein Dokument als relevant gilt (in dieser Arbeit wurde immer ein Schwellenwert von 0,5 verwendet).

Diese Arbeit beinhaltet keine Auswertung über verschiedene Prompts. Es wurden jedoch, je nachdem ob RAG genutzt wurde oder nicht, unterschiedliche Systemprompts verwendet: Anhang 3 zeigt den Systemprompt, falls RAG verwendet wird, und Anhang 4 für den Fall ohne RAG. Des Weiteren wurde auf ein Training des RAG-Systems verzichtet, es wurde das Frozen RAG-Verfahren verwendet.

Evaluationsmetriken

Um die durch die LLMs erzeugten Texte zu evaluieren, werden die Metriken Rouge [46] und BLEU [47] verwendet. Dabei erzeugt das LLM zu einem gegebenen Input einen Kandidatentext, der anschließend mit fünf Referenztexten verglichen wird. Der beste Wert, der jeweils durch Rouge und BLEU erzeugt wird, dient dann als Grundlage für die Bewertung. Beide Metriken sind darauf ausgelegt, die Qualität der automatischen Textgenerierung zu bewerten, jedoch in unterschiedlichen Bereichen: Rouge für Textzusammenfassungen und BLEU für maschinelle Übersetzungen.

Neben den Metriken Rouge und BLEU spielt auch die Dauer der Generierung eine wichtige Rolle. Hier wird insbesondere der Einfluss von RAG auf die Bearbeitungsdauer analysiert. Dies umfasst die Zeit, die benötigt wird, um relevante Dokumente zu finden und einzubeziehen. Darüber hinaus wird der durchschnittliche Stromverbrauch sowie die GPU-Auslastung gemessen, um ein umfassendes Bild der Effizienz der Modelle zu zeichnen. Eine beispielhafte Auswertung ist in Anhang 7 dargestellt.

Fragen und Referenzantworten

Da die Ausgangstexte in englischer Sprache verfasst sind, wurden alle 20 Fragen und die jeweiligen Referenzantworten, je 5 pro Frage, auf Englisch verfasst. Zur Erstellung der Fragen wurden zufällige Textpassagen des WikiText Datensatzes ausgewählt und passende Fragen-Antworten Paare formuliert. Dabei wurde nicht auf eine spezielle Verteilung oder Aufbau der Frage geachtet. Anhang 6 zeigt eine Beispielfrage mit den dazugehörigen Referenzantworten passend zu Anhang 7. Anhang 8 listet alle Fragen-Antwort Paare auf.

Rouge (Recall-Oriented Understudy for Gisting Evaluation)

Rouge ist eine Methode für die automatische Bewertung von Textzusammenfassungen, die durch den Vergleich von Computer erzeugten Zusammenfassungen mit idealen, von Menschen erstellten Zusammenfassungen deren Qualität bewertet. Es umfasst verschiedene Messmethoden wie Rouge-N, Rouge-L, Rouge-W und Rouge-S, die n-Gramm-Überschneidungen, längste gemeinsame Teilsequenzen, gewichtete längste gemeinsame Teilsequenzen und Skip-Bigram-Überschneidungen zählen. Diese Arbeit misst Rouge-1, Rouge-2, Rouge-L und Rouge-LSum. Der Rouge-Score liegt zwischen 0 und 1. Ein ROUGE-Score, auf der jeweiligen Untersuchungsebene, von 0 bedeutet keine Übereinstimmung, ein Score von 0.5 zeigt eine teilweise Übereinstimmung, und ein Score von 1 bedeutet eine vollständige Übereinstimmung

Typ	Untersuchungsebene
Rouge-1	Unigrammen (1-Gramm), Wort
Rouge-2	Bigram (2-Gramm), Wortpaare
Rouge-L	längste gemeinsamen Teilsequenz, z.B. Satzteil
Rouge-LSum	längste gemeinsamen Teilsequenz einer Zeile ("\\n" getrennt)

Tabelle 2: Rouge Typen Beschreibung

BLEU (Bilingual Evaluation Understudy)

BLEU ist eine Methode zur automatischen Bewertung der Qualität maschineller Übersetzungen. Sie zielt darauf ab, schnelle, kostengünstige und sprachunabhängige Bewertungen zu ermöglichen, die stark mit menschlichen Bewertungen korrelieren. Der BLEU-Score berechnet die Übereinstimmung der N-Gramme zwischen der maschinellen Übersetzung und den menschlichen Referenzübersetzungen, wobei ein gewichteter Durchschnitt der modifizierten N-Gramm-Genauigkeiten verwendet wird. Der BLEU-Score liegt zwischen 0 und 1. Während ein BLEU-Score von 0 bedeutet, dass die maschinelle Übersetzung keinerlei Übereinstimmung mit der Referenzübersetzung hat, deutet ein BLEU-Score von ~ 0.5 an, dass die Übersetzung teilweise mit der Referenz übereinstimmt, also von mittlerer Qualität ist. Ein BLEU-Score von 1 bedeutet eine perfekte Übereinstimmung mit der Referenzübersetzung. Höhere Werte deuten generell auf eine bessere Übereinstimmung und somit auf eine höhere Qualität der maschinellen Übersetzung hin.

Auswertung

Dieser Abschnitt der Arbeit befasst sich umfassend und detailliert mit der Bewertung verschiedener Modelle hinsichtlich ihrer Leistungsfähigkeit und Effizienz. Es werden dabei eingehende Vergleiche zwischen Modellen mit Retrieval-Augmented Generation und solchen mit Fine-Tuning vorgenommen, um deren jeweilige Stärken und Schwächen fundiert zu analysieren.

Durchgeführte Evaluationen

In Tabelle 3 sind die durchgeführten Trainings und Evaluationen der einzelnen Modelle ausführlich aufgelistet.

Modelgröße	Modellname	B	T	FT	RAG	FT u. RAG
7B - 8B	meta-llama/Meta-Llama-3-8B-Instruct	✓	X	X	✓	X
	meta-llama/Meta-Llama-3.1-8B-Instruct	✓	X	X	✓	X
	Qwen/Qwen2-7B-Instruct	✓	X	X	✓	X
	Qwen/Qwen-7B-Chat	✓	X	X	✓	X
	mistralai/Mistral-7B-Instruct-v0.3	✓	X	X	✓	X
4B	microsoft/Phi-3-mini-128k-instruct	✓	X	X	✓	X
	Qwen/Qwen1.5-4B-Chat	✓	X	X	✓	X
0B - 2B	internlm/internlm2-chat-1_8b	✓	✓	✓	✓	✓
	Qwen/Qwen1.5-1.8B-Chat	✓	✓	✓	✓	✓
	HuggingFaceTB/SmolLM-1.7B-Instruct	✓	✓	✓	✓	✓
	stabilityai/stablelm-2-1_6b-chat	✓	✓	✓	✓	✓
	Qwen/Qwen2-1.5B-Instruct	✓	✓	✓	✓	✓
	Qwen/Qwen1.5-0.5B-Chat	✓	✓	✓	✓	✓
	Qwen/Qwen2-0.5B-Instruct"	✓	✓	✓	✓	✓

Tabelle 3: Durchgeführte Trainings und Evaluationen. (T: Modell wurde trainiert; B: Basisversion des Modells, FT: Fine-Tuned-Modell; RAG: Basisversion und RAG; FT u. RAG: Fine-Tuned-Modell und RAG)

Wie bereits angesprochen, konnten aufgrund von Hardware-Einschränkungen keine Modelle mit mehr als 2 Milliarden Parametern trainiert werden. Die Evaluationen der großen Modelle dauerten teilweise sehr lange, da Parameter auf die CPU ausgelagert werden mussten.

Übersicht

Eine grobe Übersicht der Gesamtergebnisse ist in Tabelle 4 zu finden. Microsofts Phi-3-mini, ohne Fine-Tuning aber mit RAG, schneidet in allen Scores am besten ab. Das Training des Modells internlm2-chat-1_8B scheint problematisch zu sein, da es nach dem Training keinen Output mehr generiert. Bei diesem Modell wurde die Evaluation daher wiederholt, während alle anderen Modelle nur einmal evaluiert wurden.

RAG verbessert die Performance sowohl beim Basismodell als auch beim Fine-Tuned-Modell in fast allen Fällen, wobei zwischen dem Basismodell mit RAG und Fine-Tuned-Modell mit RAG auch wieder kein klarer Unterschied festzustellen ist. RAG scheint also die effektivere Methode im Vergleich zu Fine-Tuning zu sein.

Modelname	FT	RAG	BLEU Score	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Lsum
meta-llama_Meta-Llama-3.1-8B-Instruct	X	X	0.095	0.256	0.102	0.203	0.22
meta-llama_Meta-Llama-3.1-8B-Instruct	X	✓	0.261	0.453	0.316	0.422	0.43
meta-llama_Meta-Llama-3-8B-Instruct	X	X	0.146	0.291	0.142	0.258	0.262
meta-llama_Meta-Llama-3-8B-Instruct	X	✓	0.167	0.353	0.227	0.312	0.323
mistralai_Mistral-7B-Instruct-v0.3	X	X	0.109	0.252	0.113	0.212	0.219
mistralai_Mistral-7B-Instruct-v0.3	X	✓	0.235	0.393	0.284	0.364	0.366
Qwen_Qwen2-7B-Instruct	X	X	0.109	0.27	0.119	0.227	0.23
Qwen_Qwen2-7B-Instruct	X	✓	0.257	0.423	0.3	0.394	0.394
Qwen_Qwen-7B-Chat	X	X	0.109	0.242	0.104	0.2	0.206
Qwen_Qwen-7B-Chat	X	✓	0.183	0.373	0.244	0.335	0.339
Qwen_Qwen1.5-4B-Chat	X	X	0.11	0.258	0.112	0.206	0.206
Qwen_Qwen1.5-4B-Chat	X	✓	0.224	0.398	0.279	0.382	0.382
microsoft_Phi-3-mini-128k-instruct	X	X	0.163	0.308	0.156	0.28	0.282
microsoft_Phi-3-mini-128k-instruct	X	✓	0.323	0.503	0.36	0.467	0.467
internlm_internlm2-chat-1_8b	X	X	0.062	0.133	0.07	0.116	0.12
internlm_internlm2-chat-1_8b	X	✓	0.156	0.347	0.217	0.295	0.296
internlm_internlm2-chat-1_8b	✓	X	0	0	0	0	0
internlm_internlm2-chat-1_8b	✓	✓	0	0	0	0	0
Qwen_Qwen1.5-1.8B-Chat	X	X	0.014	0.067	0.023	0.052	0.058
Qwen_Qwen1.5-1.8B-Chat	X	✓	0.042	0.112	0.063	0.096	0.104
Qwen_Qwen1.5-1.8B-Chat	✓	X	0.046	0.122	0.055	0.105	0.11
Qwen_Qwen1.5-1.8B-Chat	✓	✓	0.121	0.254	0.175	0.227	0.233
HuggingFaceTB_SmolLM-1.7B-Instruct	X	X	0.073	0.21	0.084	0.172	0.175
HuggingFaceTB_SmolLM-1.7B-Instruct	X	✓	0.102	0.261	0.147	0.221	0.226
HuggingFaceTB_SmolLM-1.7B-Instruct	✓	X	0.08	0.229	0.1	0.188	0.195
HuggingFaceTB_SmolLM-1.7B-Instruct	✓	✓	0.006	0.016	0.006	0.015	0.015
stabilityai_stablelm-2-1_6b-chat	X	X	0.04	0.153	0.049	0.119	0.125
stabilityai_stablelm-2-1_6b-chat	X	✓	0.063	0.169	0.097	0.142	0.145
stabilityai_stablelm-2-1_6b-chat	✓	X	0.052	0.164	0.064	0.14	0.143
stabilityai_stablelm-2-1_6b-chat	✓	✓	0.065	0.182	0.097	0.152	0.156
Qwen_Qwen2-1.5B-Instruct	X	X	0.158	0.298	0.147	0.256	0.256
Qwen_Qwen2-1.5B-Instruct	X	✓	0.203	0.383	0.263	0.343	0.346
Qwen_Qwen2-1.5B-Instruct	✓	X	0.139	0.314	0.147	0.263	0.263
Qwen_Qwen2-1.5B-Instruct	✓	✓	0.23	0.474	0.321	0.437	0.437
Qwen_Qwen2-0.5B-Instruct	X	X	0.042	0.173	0.051	0.131	0.131
Qwen_Qwen2-0.5B-Instruct	X	✓	0.118	0.306	0.169	0.27	0.275
Qwen_Qwen2-0.5B-Instruct	✓	X	0.034	0.143	0.041	0.118	0.119
Qwen_Qwen2-0.5B-Instruct	✓	✓	0.012	0.071	0.022	0.056	0.059
Qwen_Qwen1.5-0.5B-Chat	X	X	0.077	0.207	0.076	0.176	0.18
Qwen_Qwen1.5-0.5B-Chat	X	✓	0.096	0.235	0.136	0.204	0.213
Qwen_Qwen1.5-0.5B-Chat	✓	X	0.022	0.116	0.018	0.093	0.096
Qwen_Qwen1.5-0.5B-Chat	✓	✓	0.034	0.156	0.04	0.122	0.121
Average			0.109	0.24	0.132	0.209	0.212

Tabelle 4: Modellübersicht BLEU und Rouge Scores (FT: Fine-Tuning-Model, RAG: mit RAG System). Sortiert nach Modellgröße: groß nach klein.

Zweistichproben-t-Test

Die nachfolgende Tabelle 5 zeigt die Ergebnisse des T-Tests der Varianten FT, RAG und FT + RAG gegenüber dem Basismodell ohne FT und ohne RAG. Der T-Wert gibt den Unterschied zwischen den Mittelwerten der zwei Stichproben relativ zur Streuung der Daten an. Eine hohe absolute T-Statistik

deutet darauf hin, dass der Unterschied zwischen den Stichprobenmittelwerten nicht zufällig ist. Der p-Wert gibt die Wahrscheinlichkeit an, dass der beobachtete Unterschied zwischen den Stichprobenmittelwerten unter der Annahme der Nullhypothese auftritt. In diesem Fall ist die Nullhypothese, dass FT oder RAG keinen Unterschied machen. Ein p-Wert unter 0.05 gibt typischerweise an, dass die Nullhypothese abgelehnt werden kann und der beobachtete Unterschied statistisch signifikant ist [48].

Die Effektstärke wird hier in Farbkategorien angegeben:

- Weiß entspricht einem schwachen Effekt mit $r < 0.3$
- Gelb entspricht einem mittleren Effekt mit $0.3 \leq r < 0.5$
- Grün stellt einen starken Effekt mit $r \geq 0.5$ dar.

Folgende Gleichungen werden zur Berechnung des Korrelationskoeffizienten r benötigt:

$$t = \frac{\text{Mittelwert(Evaluation Basis)} - \text{Mittelwert(Evaluation Variante)}}{\text{Standardfehler der Mittelwertsunterschiede}}$$

$$df = \text{Stichprobengröße(Evaluation Basis)} + \text{Stichprobengröße(Evaluation Variante)} - 2$$

$$r = \sqrt{t^2 \div (t^2 + df)}$$

Negative T-Werte, Resultat daraus, dass der jeweilige Score Mittelwert der Variante größer ist als der Mittelwerte des Scores aus der Basisevaluation, stellen eine positive Veränderung der Variante gegenüber dem Basismodell dar, das Modell liefert bessere Ausgaben. Es ist erkennbar, dass vor allem die größeren Modelle häufig von RAG profitieren. Auch die kleineren Modelle zeigen stellenweise signifikante Veränderungen, allerdings sind ihre Verbesserungen nicht so stark ausgeprägt. Die T-Werte verbessern sich häufig, zeigen insgesamt aber eher durchwachsene Verbesserungen. Internlm2-chat-1_8B zeigt den erwarteten starken Effekt; nach FT wurde kein Text mehr generiert, was zu einem schlechten Score führt. Auch Gwenn2-0.5B weist eine signifikante Verschlechterung der Scores bei Verwendung von FT und RAG auf, 6 von 20 Fragen erhalten keine Antwort. RAG wirkt stark positiv auf Qwen1.5-1.8B und Microsofts Phi-3-mini-Modell aus.

Model Name	FT	RAG	BLEU	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Lsum
meta-llama_Meta-Llama-3.1-8B-Instruct	X	✓	-2.449 (0.021) [0.369]	-2.843 (0.007) [0.419]	-3.065 (0.005) [0.445]	-3.154 (0.003) [0.455]	-3.053 (0.004) [0.444]
meta-llama_Meta-Llama-3-8B-Instruct	X	✓	-0.385 (0.702) [0.062]	-0.947 (0.350) [0.152]	-1.555 (0.128) [0.245]	-0.872 (0.389) [0.140]	-0.988 (0.330) [0.158]
mistralai_Mistral-7B-Instruct-v0.3	X	✓	-2.078 (0.045) [0.319]	-2.515 (0.016) [0.378]	-2.865 (0.007) [0.421]	-2.636 (0.012) [0.393]	-2.583 (0.014) [0.387]
Qwen_Qwen-7B-Chat	X	✓	-1.314 (0.197) [0.209]	-2.396 (0.022) [0.362]	-2.505 (0.017) [0.376]	-2.443 (0.019) [0.368]	-2.415 (0.021) [0.365]
Qwen_Qwen2-7B-Instruct	X	✓	-2.502 (0.017) [0.376]	-2.601 (0.013) [0.389]	-3.079 (0.004) [0.447]	-2.770 (0.009) [0.410]	-2.734 (0.009) [0.405]
Qwen_Qwen1.5-4B-Chat	X	✓	-2.101 (0.043) [0.323]	-2.296 (0.027) [0.349]	-2.805 (0.008) [0.414]	-2.881 (0.007) [0.423]	-2.868 (0.007) [0.422]
microsoft_Phi-3-mini-128k-instruct	X	✓	-2.490 (0.018) [0.375]	-3.135 (0.003) [0.453]	-3.591 (0.001) [0.503]	-2.972 (0.005) [0.434]	-2.955 (0.005) [0.432]
internlm_internlm2-chat-1_8b	X	✓	-2.264 (0.029) [0.345]	-4.468 (0.000) [0.587]	-2.972 (0.005) [0.434]	-3.704 (0.001) [0.515]	-3.631 (0.001) [0.508]
internlm_internlm2-chat-1_8b	✓	X	2.141 (0.045) [0.328]	4.102 (0.001) [0.554]	2.187 (0.041) [0.334]	3.602 (0.002) [0.505]	3.781 (0.001) [0.523]
internlm_internlm2-chat-1_8b	✓	✓	2.141 (0.045) [0.328]	4.102 (0.001) [0.554]	2.187 (0.041) [0.334]	3.602 (0.002) [0.505]	3.781 (0.001) [0.523]
Qwen_Qwen1.5-1.8B-Chat	X	✓	-3.376 (0.003) [0.480]	-3.829 (0.001) [0.528]	-4.179 (0.000) [0.561]	-4.021 (0.000) [0.546]	-4.038 (0.000) [0.548]
Qwen_Qwen1.5-1.8B-Chat	✓	X	-1.068 (0.299) [0.171]	-1.819 (0.083) [0.283]	-1.087 (0.290) [0.174]	-1.750 (0.096) [0.273]	-1.718 (0.101) [0.268]
Qwen_Qwen1.5-1.8B-Chat	✓	✓	-1.963 (0.064) [0.303]	-3.281 (0.004) [0.470]	-2.457 (0.024) [0.370]	-2.972 (0.008) [0.434]	-2.970 (0.008) [0.434]
HuggingFaceTB_SmolLM-1.7B-Instruct	X	✓	-0.769 (0.447) [0.124]	-0.938 (0.355) [0.150]	-1.492 (0.145) [0.235]	-0.887 (0.381) [0.142]	-0.930 (0.358) [0.149]
HuggingFaceTB_SmolLM-1.7B-Instruct	✓	X	-0.197 (0.845) [0.042]	-0.381 (0.706) [0.062]	-0.452 (0.654) [0.073]	-0.341 (0.735) [0.055]	-0.437 (0.664) [0.071]
HuggingFaceTB_SmolLM-1.7B-Instruct	✓	✓	3.037 (0.006) [0.432]	5.084 (0.000) [0.636]	3.301 (0.003) [0.472]	4.256 (0.000) [0.568]	4.356 (0.000) [0.577]
stabilityai_stablelm-2-1_6b-chat	X	✓	-1.503 (0.141) [0.237]	-0.590 (0.559) [0.095]	-2.402 (0.021) [0.363]	-1.037 (0.306) [0.166]	-0.881 (0.384) [0.141]
stabilityai_stablelm-2-1_6b-chat	✓	X	-0.654 (0.518) [0.106]	-0.373 (0.711) [0.060]	-0.694 (0.492) [0.112]	-0.808 (0.424) [0.130]	-0.682 (0.500) [0.110]
stabilityai_stablelm-2-1_6b-chat	✓	✓	-1.176 (0.249) [0.187]	-0.914 (0.367) [0.147]	-2.049 (0.048) [0.315]	-1.250 (0.219) [0.199]	-1.154 (0.256) [0.184]
Qwen_Qwen2-1.5B-Instruct	X	✓	-0.704 (0.486) [0.113]	-1.401 (0.170) [0.222]	-1.815 (0.078) [0.283]	-1.364 (0.181) [0.216]	-1.409 (0.167) [0.223]
Qwen_Qwen2-1.5B-Instruct	✓	X	0.305 (0.762) [0.049]	-0.289 (0.774) [0.047]	0.003 (0.997) [0.001]	-0.117 (0.908) [0.019]	-0.119 (0.906) [0.019]
Qwen_Qwen2-1.5B-Instruct	✓	✓	-1.208 (0.235) [0.192]	-2.912 (0.006) [0.427]	-2.643 (0.012) [0.394]	-2.803 (0.008) [0.414]	-2.816 (0.008) [0.416]
Qwen_Qwen1.5-0.5B-Chat	X	✓	-0.626 (0.535) [0.101]	-0.703 (0.486) [0.113]	-1.796 (0.081) [0.280]	-0.691 (0.494) [0.111]	-0.807 (0.425) [0.130]
Qwen_Qwen1.5-0.5B-Chat	✓	X	2.278 (0.033) [0.347]	2.642 (0.013) [0.394]	2.600 (0.017) [0.389]	2.612 (0.015) [0.390]	2.661 (0.013) [0.396]
Qwen_Qwen1.5-0.5B-Chat	✓	✓	1.708 (0.101) [0.267]	1.304 (0.201) [0.207]	1.351 (0.186) [0.214]	1.549 (0.131) [0.244]	1.720 (0.095) [0.269]
Qwen_Qwen2-0.5B-Instruct	X	✓	-3.360 (0.003) [0.479]	-3.363 (0.002) [0.479]	-3.310 (0.003) [0.473]	-3.885 (0.001) [0.533]	-4.092 (0.000) [0.553]
Qwen_Qwen2-0.5B-Instruct	✓	X	0.673 (0.506) [0.108]	1.041 (0.305) [0.166]	0.632 (0.531) [0.102]	0.587 (0.561) [0.095]	0.532 (0.598) [0.086]
Qwen_Qwen2-0.5B-Instruct	✓	✓	4.082 (0.000) [0.552]	3.692 (0.001) [0.514]	2.244 (0.033) [0.342]	3.714 (0.001) [0.516]	3.547 (0.001) [0.499]

Tabelle 5: Werte: T-Statistik t (Signifikanzwert p) [Effektstärke r], Farbcode: Weiß=schwacher Effekt ($r < 0.3$), Gelb=mittlerer Effekt ($r < 0.5$), Grün=starker Effekt ($r \geq 0.5$)

Durchschnittliche BLEU und Rouge Scores

Die Abbildungen 2 und 3 veranschaulichen die durchschnittlichen BLEU- und Rouge1-Scores der verschiedenen Modelle. Es ist jedoch wichtig zu beachten, dass Modelle mit mehr als 2 Milliarden Parametern, die keinem Fine-Tuning unterzogen werden konnten, in diesen Darstellungen fehlen.

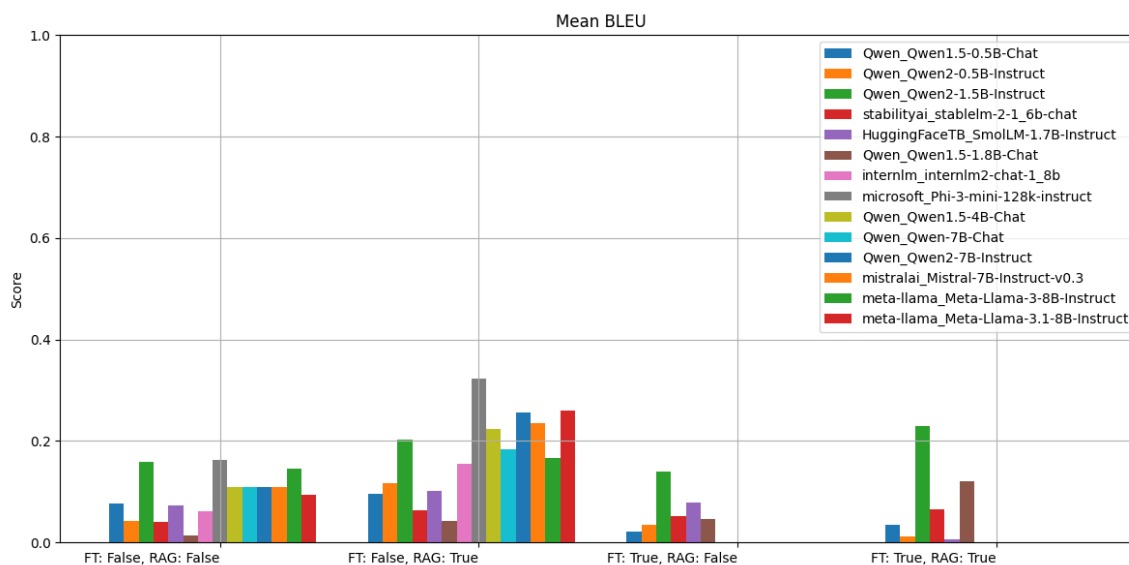


Abbildung 2: Modellübersicht durchschnittliche BLEU Scores

Diese Darstellungen veranschaulichen sehr gut, dass ein Basismodell in Kombination mit RAG effektiver ist als das Basismodell allein, ein Fine-Tuned-Modell sowie Fine-Tuned-Modelle in Kombination mit RAG. Das Basismodell erhält meist bessere Bewertungen als das FT-Modell. Es wäre darüber hinaus äußerst interessant zu untersuchen, wie Modelle mit mehr als zwei Milliarden Parametern auf das Fine-Tuning reagieren. Microsofts Phi-3-Modell schneidet in den beiden hier gezeigten Scores am besten ab.

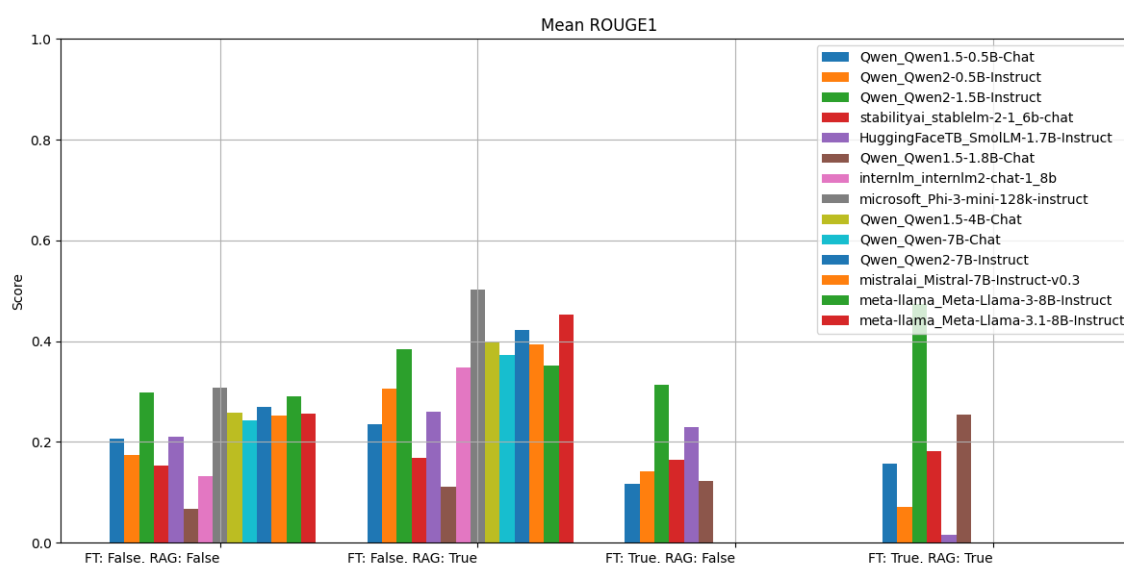


Abbildung 3: Modellübersicht durchschnittliche Rouge-1 Scores

Im Anhang ist eine Übersicht aller Scores zu finden.

Score-Verteilung

Die Abbildungen 4 und 5 illustrieren die Verteilung der Bleu- und Rouge-Scores für die 20 gestellten Fragen. Im Anhang 15 ist zudem eine Gegenüberstellung aller Scores zu finden. Ein Boxplot ist ein Diagramm zur Darstellung der Verteilung eines Datensatzes. Es zeigt das Minimum, das untere Quartil, den Median, das obere Quartil und das Maximum. Der Strich in der Box repräsentiert den Median, also den mittleren Wert des Datensatzes. Zusätzlich werden Ausreißer angezeigt.

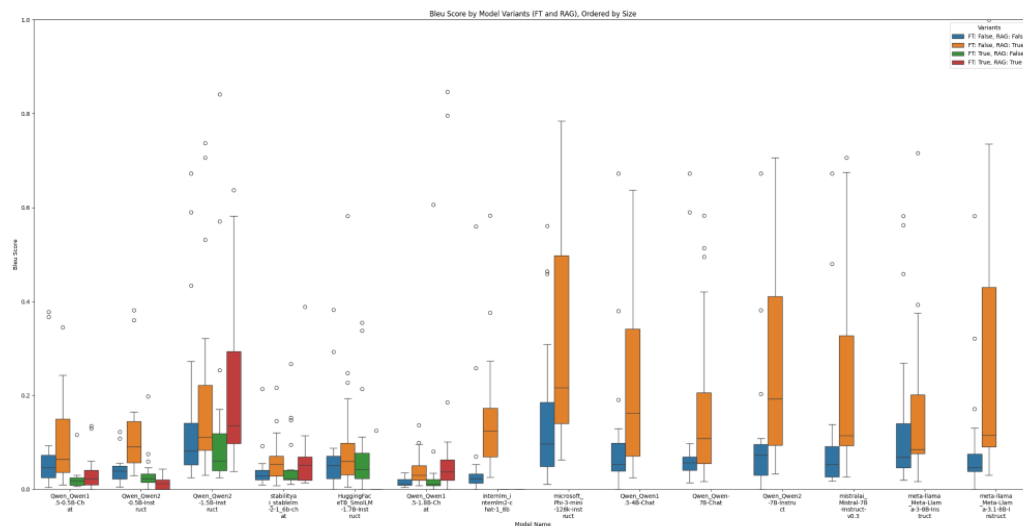


Abbildung 4: Boxplot BLEU Score Distribution

Das Modell Qwen2-1.5B-Instruct schneidet im Vergleich zu seinen Kontrahenten derselben Modellgröße besonders gut ab und erzielt in der Variante Fine-Tuning plus Retrieval-Augmented Generation sogar bessere Rouge-1-Ergebnisse als wesentlich größere Modelle in der Basisversion oder in der Kombination Basis plus RAG. Sowohl Phi-3-mini als auch Llama-3_1-8B scheinen darüber hinaus ebenfalls gute Kandidaten zu sein.

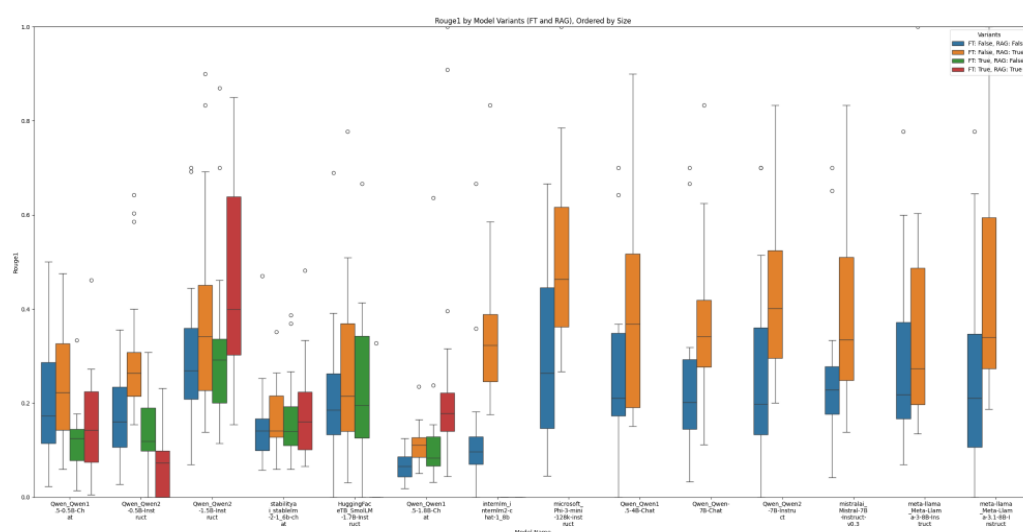


Abbildung 5: Boxplot Rouge1 Score Distribution

Bearbeitungsdauer

Bei der Untersuchung der Bearbeitungsdauer haben sich zwei klar unterschiedliche Gruppen gebildet: die „schnelle“ Gruppe, die in Abbildung 6 dargestellt ist, und die „langsame“ Gruppe, zu sehen in Abbildung 7. Zu beachten ist, dass die schnelle Gruppe maximal etwa 40 Sekunden benötigte, im Durchschnitt nur etwa 10 Sekunden pro Frage. Im Gegensatz dazu benötigte die langsame Gruppe in der Regel mindestens 50 Sekunden pro Frage. Eine mögliche Erklärung für diese Differenz könnte darin liegen, dass die Modelle der langsamen Gruppe zu groß für die verfügbare GPU waren und daher teilweise ausgelagert werden mussten.

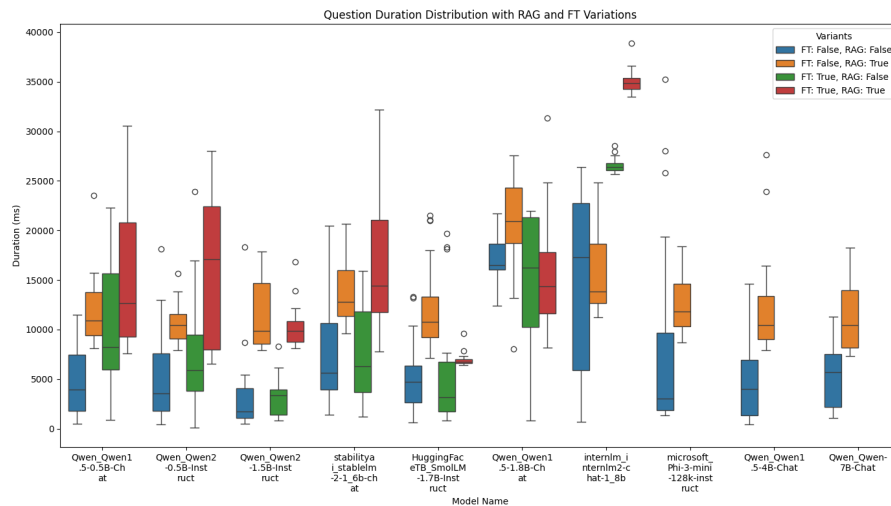


Abbildung 6: Boxplot Bearbeitungsdauer der Fragen in der „schnellen“ Gruppe

Interessant ist, dass Modelle wie phi-3-mini, Qwen1.5-4B und Qwen-7B in der schnellen Gruppe zu finden sind. Diese Modelle sind teilweise sogar schneller als einige der kleineren Modelle. Der deutliche Unterschied lässt den Schluss zu, dass auch größere Modelle in der Lage sind, schnell zu antworten. Es wurde sich gegen eine logarithmische Darstellung entschieden, da die Werte in der schnellen Gruppe sonst nicht mehr unterscheidbar gewesen wären.

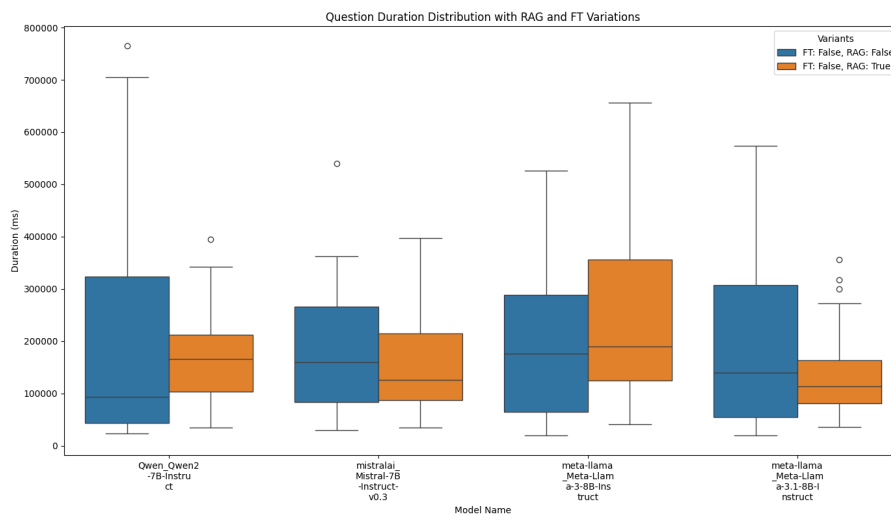


Abbildung 7: Boxplot Bearbeitungsdauer der Fragen in der „langsamen“ Gruppe

Modellgröße im Verhältnis zu BLEU und Rouge

Es zeigt sich der Trend, dass größere Modelle tendenziell bessere Ergebnisse liefern, wie in Abbildung 8 und Abbildung 9 zu sehen ist. Dabei ist darauf zu achten, dass die y-Achse in diesen Graphen nicht wie zuvor von 0 bis 1 skaliert ist, um unnötigen Leerraum zu vermeiden. Die Größe der Datenpunkte wird durch den durchschnittlichen Energieverbrauch der Modelle bestimmt.

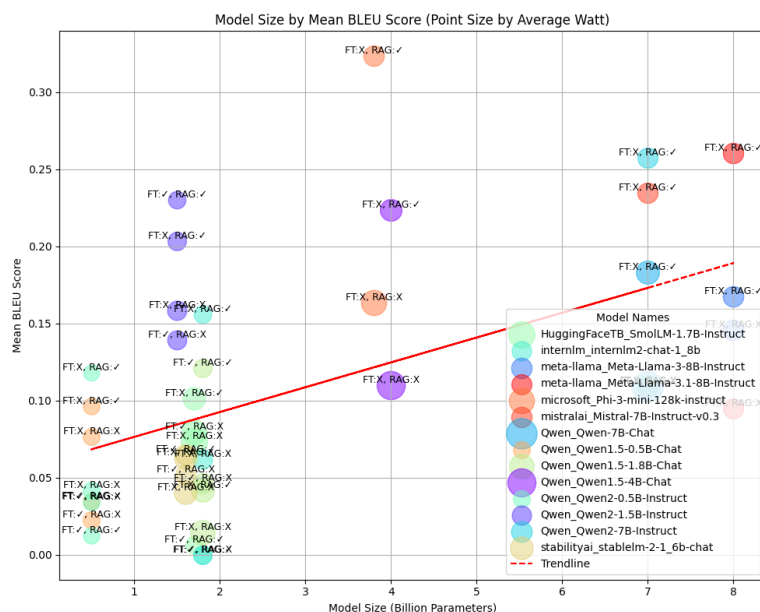


Abbildung 8: Modellgröße im Verhältnis zum durchschnittlichen BLEU-Score mit Trendlinie und Energieverbrauch

In dieser Grafik erweisen sich insbesondere die Modelle Qwen2-1.5B und Phi-3-mini als überdurchschnittlich leistungsfähig für ihre jeweilige Modellgröße. Es lässt sich ein deutlicher Aufwärtstrend feststellen, sobald größere Modelle verwendet werden.

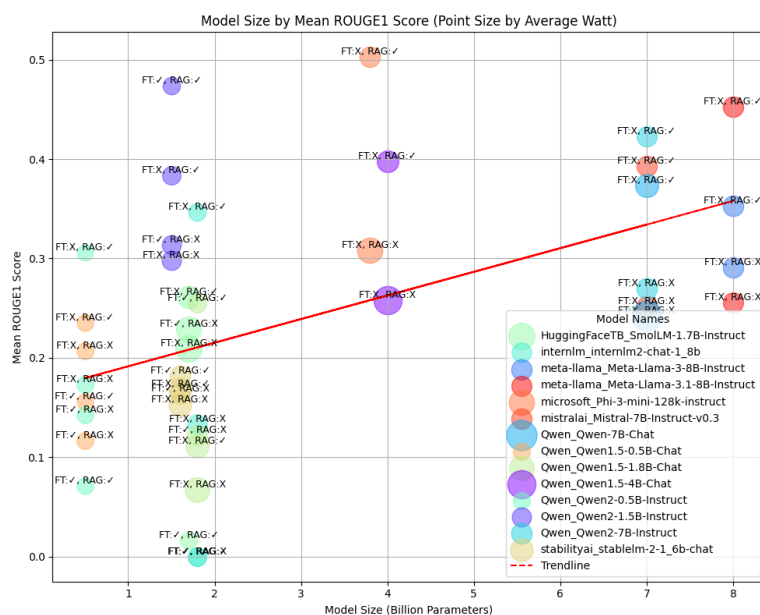


Abbildung 9: Modellgröße im Verhältnis zum durchschnittlichen Rouge-1-Score mit Trendlinie und Energieverbrauch

GPU-Metriken: Auslastungsverlauf und Unterschiede mit FT und RAG

In den folgenden Abbildungen 10 und 11 wird der typische Verlauf der GPU-Auslastung während der Textgenerierung visualisiert. Die Diagramme auf der linken Seite stellen jeweils die Auslastung ohne Retrieval-Augmented Generation dar, während die rechten Diagramme die Auslastung mit RAG darstellen.

GPU Metrics for Qwen1.5-4B-Chat - Question 4

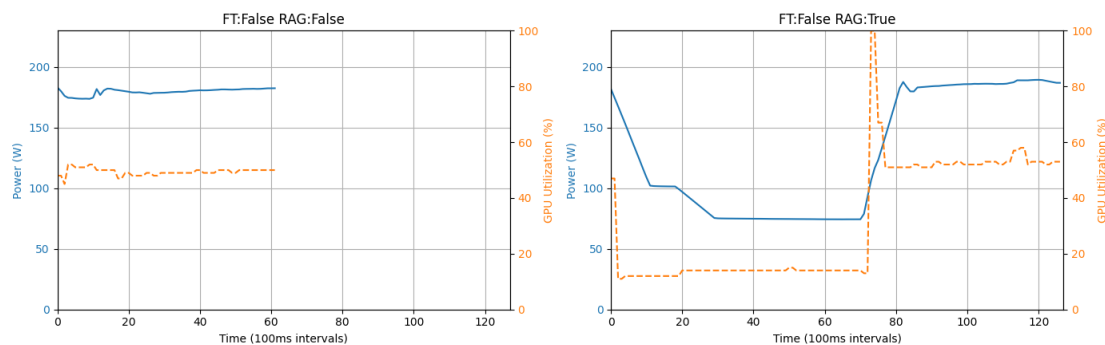


Abbildung 10: GPU-Metrik für Modell Qwen1.5-4B bei Frage 4

Es ist deutlich zu erkennen, dass RAG, sichtbar durch die charakteristischen „Senken“ in den Graphen, einen nicht zu vernachlässigenden Anteil an der Verarbeitungszeit hat. Zunächst müssen geeignete Dokumente gefunden werden. Im Anschluss daran benötigen die LLMs wahrscheinlich länger, da sie einen größeren Input verarbeiten müssen.

GPU Metrics for Qwen2-1.5B-Instruct - Question 4

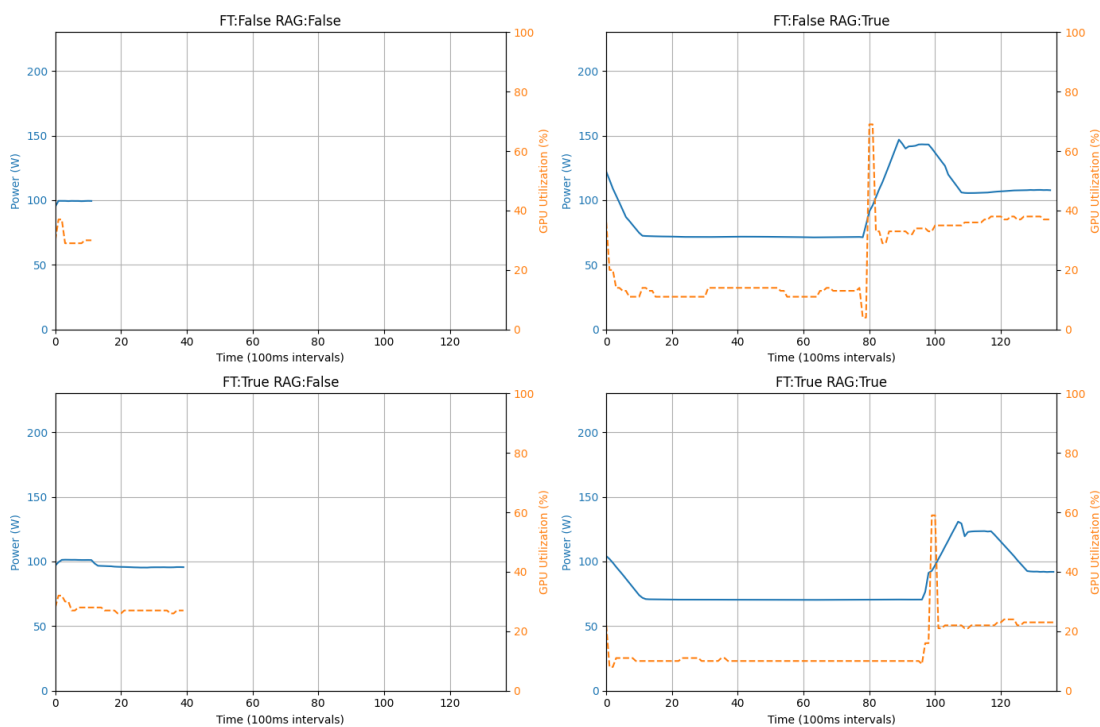


Abbildung 11: GPU-Metrik für Modell Qwen2-1.5B bei Frage 4

Tabelle 5 illustriert die Dauer sowie die durchschnittliche GPU-Auslastung bei der Evaluation der jeweiligen Modelle und ihrer entsprechenden Varianten. Dabei ist auffällig, dass große Modelle tendenziell einen ähnlichen, wenn auch leicht erhöhten Energieverbrauch im Vergleich zu den kleineren Modellen zeigen. Diese Beobachtung lässt sich besonders auf die intensivere Auslastung der GPU zurückführen, die bei größeren Modellen höher ist. Durch die längere Evaluationsdauer steigen jedoch der Gesamtenergieverbrauch und die Beanspruchung der Ressourcen, was letztlich dazu führt, dass große Modelle insgesamt mehr Energie verbrauchen.

Modelname	FT	RAG	Total Duration (min:sec)	Avg Power (W)	Avg Utilization (%)	Total Consumed Power (Wh)
meta-llama_Meta-Llama-3.1-8B-Instruct	X	X	67:32	107,15	51,75	120,63
meta-llama_Meta-Llama-3.1-8B-Instruct	X	✓	47:29	110,23	54,61	87,24
meta-llama_Meta-Llama-3-8B-Instruct	X	X	68:30	110,73	54,67	126,45
meta-llama_Meta-Llama-3-8B-Instruct	X	✓	82:14	111,59	57,15	152,97
mistralai_Mistral-7B-Instruct-v0.3	X	X	63:43	110,50	55,53	117,36
mistralai_Mistral-7B-Instruct-v0.3	X	✓	55:59	109,40	50,85	102,08
Qwen_Qwen2-7B-Instruct	X	X	75:38	112,36	65,06	141,64
Qwen_Qwen2-7B-Instruct	X	✓	57:27	107,68	53,84	103,13
Qwen_Qwen-7B-Chat	X	X	1:54	208,96	48,88	6,62
Qwen_Qwen-7B-Chat	X	✓	3:44	134,36	30,02	8,38
Qwen_Qwen1.5-4B-Chat	X	X	1:37	184,27	48,55	4,98
Qwen_Qwen1.5-4B-Chat	X	✓	4:08	121,34	28,70	8,37
microsoft_Phi-3-mini-128k-instruct	X	X	2:50	154,19	37,75	7,30
microsoft_Phi-3-mini-128k-instruct	X	✓	4:15	110,70	23,27	7,87
internlm_internlm2-chat-1_8b	X	X	4:47	102,25	28,76	8,18
internlm_internlm2-chat-1_8b	X	✓	5:10	87,87	16,95	7,58
internlm_internlm2-chat-1_8b	✓	X	8:51	93,93	27,75	13,87
internlm_internlm2-chat-1_8b	✓	✓	11:41	92,72	25,53	18,08
Qwen_Qwen1.5-1.8B-Chat	X	X	5:44	148,26	38,43	14,19
Qwen_Qwen1.5-1.8B-Chat	X	✓	6:48	133,24	31,92	15,13
Qwen_Qwen1.5-1.8B-Chat	✓	X	4:58	107,97	33,55	8,95
Qwen_Qwen1.5-1.8B-Chat	✓	✓	5:04	94,54	23,75	8,00
HuggingFaceTB_SmolLM-1.7B-Instruct	X	X	1:41	165,49	46,88	4,69
HuggingFaceTB_SmolLM-1.7B-Instruct	X	✓	4:04	127,12	30,35	8,63
HuggingFaceTB_SmolLM-1.7B-Instruct	✓	X	1:56	150,74	37,89	4,88
HuggingFaceTB_SmolLM-1.7B-Instruct	✓	✓	2:19	85,22	16,81	3,30
stabilityai_stablelm-2-1_6b-chat	X	X	2:32	131,11	35,08	5,56
stabilityai_stablelm-2-1_6b-chat	X	✓	4:33	132,08	30,78	10,04
stabilityai_stablelm-2-1_6b-chat	✓	X	2:32	107,95	33,90	4,57
stabilityai_stablelm-2-1_6b-chat	✓	✓	5:17	110,98	26,71	9,80
Qwen_Qwen2-1.5B-Instruct	X	X	1:07	100,61	30,45	1,88
Qwen_Qwen2-1.5B-Instruct	X	✓	3:55	93,45	21,89	6,11
Qwen_Qwen2-1.5B-Instruct	✓	X	1:05	98,66	27,37	1,78
Qwen_Qwen2-1.5B-Instruct	✓	✓	3:25	85,59	17,49	4,90
Qwen_Qwen2-0.5B-Instruct	X	X	1:49	81,82	18,43	2,50
Qwen_Qwen2-0.5B-Instruct	X	✓	3:29	74,71	16,55	4,34
Qwen_Qwen2-0.5B-Instruct	✓	X	2:26	80,08	16,13	3,26
Qwen_Qwen2-0.5B-Instruct	✓	✓	5:25	77,40	16,93	7,00
Qwen_Qwen1.5-0.5B-Chat	X	X	1:33	81,47	22,27	2,11
Qwen_Qwen1.5-0.5B-Chat	X	✓	3:59	78,83	18,81	5,25
Qwen_Qwen1.5-0.5B-Chat	✓	X	3:32	81,82	19,07	4,82
Qwen_Qwen1.5-0.5B-Chat	✓	✓	5:16	77,50	17,87	6,82
Average			15:31	111,35	33,07	28,36

Tabelle 6: Modellübersicht Bearbeitungsdauer und GPU Metriken des Tests (FT: Fine-Tuning-Model, RAG: mit RAG System). Sortiert nach Modellgröße: groß nach klein.

Vergleich zwischen Dauer und ROUGE-1-Score

Im Folgenden wird Abbildung 12 vorgestellt, die einen Vergleich zwischen der Antwortzeit (logarithmisch dargestellt) und der ROUGE-1-Bewertung zeigt. Einige wenige Datenpunkte wurden entfernt, um den Anzeigebereich des Graphen kleiner zu gestalten. Die angezeigte Trendlinie verläuft durch das Streudiagramm, ist jedoch relativ flach, was auf eine minimale Korrelation zwischen der Dauer und dem ROUGE-1-Score hinweist. Dies lässt darauf schließen, dass eine längere

Antwortdauer nicht zwangsläufig mit einer Verbesserung oder Verschlechterung der ROUGE-1-Punktzahl verbunden ist.

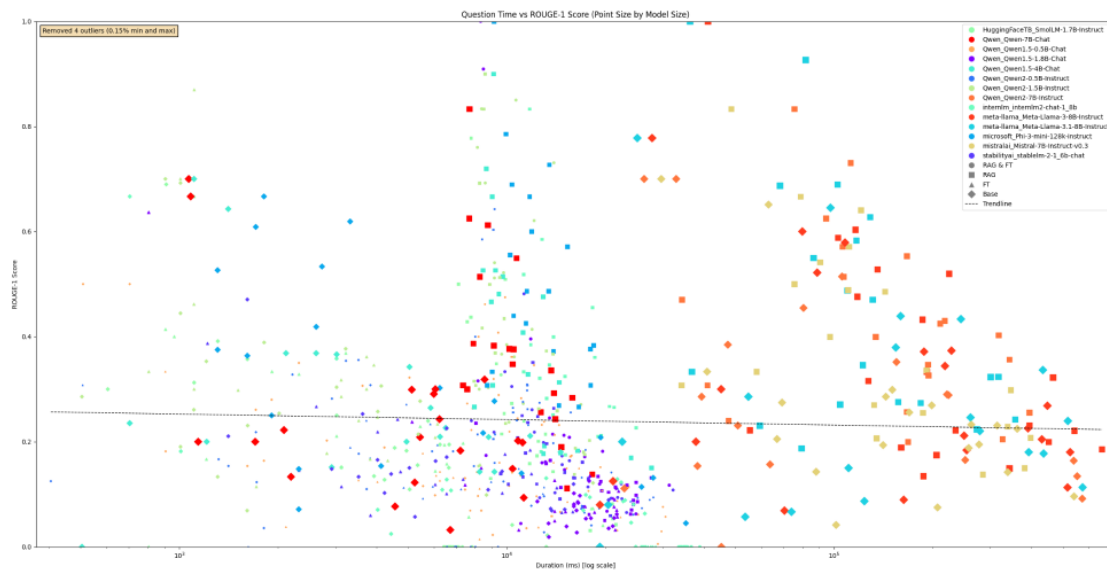


Abbildung 12: Vergleich der Antwortzeit mit dem ROUGE-1-Score

Ausgabebetrachtung

Nach der Betrachtung von reinen Zahlen wird, exemplarisch an zwei Beispielen, auf Probleme bei der alleinigen Betrachtung dieser Metriken aufmerksam gemacht.

Das erste Beispiel betrifft das Modell Qwen/Qwen-7B-Chat und Retrieval-Augmented Generation. Bei der Frage „What ship did Markgraf fire on during the Battle of Jutland?“ liefert das Modell die Antwort „The ship that Markgraf fired on during the Battle of Jutland was not mentioned in the provided documents.“ Diese Antwort erhält in allen Scores überdurchschnittliche Bewertungen (BLEU Score: ~ 0.201, Rouge-1: ~0.387, Rouge-2: ~0.206, Rouge-L: ~0.322, Rouge-Lsum: ~0.322). Sie besagt jedoch lediglich, dass keine relevanten Informationen in den bereitgestellten Dokumenten gefunden wurden.

Bei genauerer Betrachtung der durch RAG bereitgestellten Dokumente stellt sich heraus, dass diese Aussage korrekt ist. Der Wikitext-Datensatz enthält jedoch auch relevante Textabschnitte wie „Markgraf opened fire on the battlecruiser Tiger at ...“ und „Markgraf was present during ... the Battle of Jutland“, die jedoch nicht ausgewählt wurden. Dies könnte auf eine Limitierung der Anzahl der durch RAG zur Verfügung gestellten Dokumente zurückzuführen sein. Andere LLMs haben die gleichen nicht kontext relevanten Dokumente erhalten, konnten jedoch durch Vorwissen korrekt antworten. Eine Prüfung von Fine-Tuning gegenüber RAG könnte angebracht sein, jedoch ohne die tatsächliche Suche durch das RAG-System und das Bereitstellen händisch ausgewählter relevanter Dokumente.

Das zweite Beispiel zeigt, wie durch Raten ebenfalls gute Ergebnisse erzielt werden können. In der Frage „When were the Romanian Land Forces founded?“ liefert das Modell Qwen/Qwen2-1.5B-Instruct Antworten wie „The Romanian Land Forces were founded on March 12, 1948.“ und „The Romanian Land Forces were founded on October 12, 1940.“ Diese Antworten sind

zwar inkorrekt, erhalten jedoch herausragende BLEU- und vor allem Rouge Scores. Es ist somit ratsam, Fragen bedacht zu designen und auszuwählen.

Diskussion

Anders als ursprünglich angenommen, bietet die Kombination aus Retrieval-Augmented Generation und Fine-Tuning keinen so großen Vorteil wie erwartet. Verschiedene Modelle reagieren unterschiedlich auf das Fine-Tuning, sogar soweit, dass sie in einigen Fällen keine Ausgabe mehr liefern. Hier sind weitere Evaluationen notwendig, da ein Fine-Tuning der LLMs wesentlich komplexer als RAG ist. Jedes Modell zeigte jedoch eine positive Reaktion auf zusätzliche Informationen und konnte seine Ausgabe dadurch verbessern. Wie vermutet, erzeugen größere Modelle bessere Ergebnisse, allerdings ist ihr Stromverbrauch entsprechend erhöht. Dennoch bleibt die Bearbeitungsdauer ungefähr gleich. Die Antwortqualität variiert stark innerhalb eines Modells und seiner Varianten. Eine sorgfältige Auswahl der Fragen ist von großer Bedeutung, damit die Scores aussagekräftige Werte liefern.

Für den verwendeten Datensatz und die untersuchten Modelle sind folgende Modelle die beste Wahl: In der Kategorie bis 2 Milliarden Parameter ist das Modell Qwen/Qwen2-1.5B-Instruct empfehlenswert. In der Kategorie von 2 bis 4 Milliarden Parametern ist das Modell microsoft/Phi-3-mini-128k-instruct die beste Wahl. In der Kategorie über 4 Milliarden Parameter sind die Modelle Qwen/Qwen2-7B-Instruct oder meta-llama/Meta-Llama-3.1-8B-Instruct am besten geeignet.

Einschränkungen

Aufgrund der begrenzten GPU-Kapazität konnten Modelle mit mehr als 2 Milliarden Parametern nicht trainiert werden. Zudem wurden bei der Evaluation Parameter auf die GPU ausgelagert, was die Bearbeitungsdauer stark verlängert hat. Modelle, die mehr als 8 Milliarden Parameter umfassen, konnten nicht in die Bewertung aufgenommen werden.

Alle in dieser Arbeit genutzten Texte, Fragen, Referenzantworten und generierten Texte sind auf Englisch verfasst. Es konnte gezeigt werden, dass LLMs je nach Sprache unterschiedlich gute Ergebnisse liefern. Insbesondere englische und deutsche Texte sowie andere Sprachen, die häufig für das Training von LLMs verwendet werden, schneiden besser ab als Sprachen, die nur selten zum Training beitragen [49]. Daher ist es möglich, dass die Ergebnisse dieser Arbeit auch bei der Verwendung der deutschen Sprache ähnlich ausfallen könnten.

Es wurde sich im Kontext von RAG nicht mit dem sogenannten Chunking beschäftigt, also der Frage, wie Dokumente in einzelne Fragmente unterteilt werden. Eine auf die Problemstellung angepasste Strategie ist für gute Ergebnisse unerlässlich. In dieser Arbeit wurde die vorgegebene Aufteilung innerhalb des WikiText-Datensatzes eingehalten.

Wie bereits in der Auswertung erwähnt, ist die Betrachtung der einzelnen Scores nicht immer vollkommen aussagekräftig. Die Scores können eine Tendenz erkennen lassen, welche Modelle möglicherweise wert sind, näher betrachtet zu werden. Letztendlich müssen aber auch die von den Modellen erzeugten Texte gesichtet werden, um einen umfassenden Eindruck von der Qualität der Ausgabe zu gewinnen.

Zukünftige Arbeiten

Um ein umfassenderes Bild zu erhalten, sollten zukünftige Untersuchungen Modelle mit mehr als 8 Milliarden Parametern einbeziehen. Zudem ist es sinnvoll, auch kommerzielle LLM-Angebote in Betracht zu ziehen, die über APIs zugänglich sind. Es gibt mittlerweile eine Vielzahl an Anbietern, die leistungsfähige LLMs zur Verfügung stellen.

Das Fine-Tuning erwies sich leider auch bei kleineren Modellen als weniger erfolgreich als erhofft. Es sollten daher alternative Parameter für das Training untersucht werden. Im Bereich des maschinellen Lernens wird häufig von Training und dem Problem des Overfitting berichtet. Es stellt sich die Frage, ob dieses Phänomen auch bei speziellen, abgeschlossenen Datensätzen auftritt oder ob es nur eine Verallgemeinerung einer spezifischen Aufgabe behindert.

In dieser Arbeit wurden ausschließlich Instruk-Fine-Tuned-Modelle evaluiert. Es wäre interessant zu untersuchen, ob die Reihenfolge der Trainingsphasen einen großen Unterschied macht. Möglicherweise sollten LLMs nach diesem Schema trainiert werden: Foundation Model, domänenspezifisches Modell, Instruk-Finetuning um bessere Ergebnisse zu erzielen.

Die Betrachtung der Textausgaben von LLMs wirft Fragen auf, ob und wie Fragen im Evaluationskontext gestellt werden sollten. Gibt es beispielsweise für den Wikitext-Datensatz hochwertige Standardfragen?

Wie in der Einleitung erwähnt, ist diese neue Art der Interaktion mit Wissensdatenbanken spannend. Nach Abschluss dieser Arbeit beabsichtige ich, dies in einem praktischen Selbstversuch in meinem beruflichen Umfeld anzuwenden.

Zusammenfassung

Die Ergebnisse deuten darauf hin, dass RAG in Kombination mit Basismodellen besonders effektiv ist. Fine-Tuning allein zeigt geringeren Erfolg als erwartet, während größere Modelle allgemein bessere Ergebnisse liefern. Diese Erkenntnisse sind bedeutend für zukünftige Entwicklungen und Anwendungen von LLMs in dynamischen Umgebungen, in denen aktuelle und genaue Daten unerlässlich sind. Schließlich werden Einschränkungen der Studie sowie Empfehlungen für zukünftige Forschungen dargelegt, um ein umfassenderes Verständnis der Leistungsfähigkeit und Effizienz dieser Methoden zu erreichen. Die Evaluationsergebnisse und der Programmcode zum Trainieren und Evaluieren der LLMs, sowie zum Visualisieren der Ergebnisse sind in einem Git Repository im Fokus Fraunhofer GitLab zusammengefasst und können dort eingesehen werden. [50]

Quellenverzeichnis

- [1] R. Bommasani u. a., „On the Opportunities and Risks of Foundation Models“, 12. Juli 2022, *arXiv*: arXiv:2108.07258. doi: 10.48550/arXiv.2108.07258.
- [2] A. Shahid und M. Mushtaq, „A Survey Comparing Specialized Hardware And Evolution In TPUs For Neural Networks“, in *2020 IEEE 23rd International Multitopic Conference (INMIC)*, Nov. 2020, S. 1–6. doi: 10.1109/INMIC50486.2020.9318136.
- [3] V. Rawte, A. Sheth, und A. Das, „A Survey of Hallucination in Large Foundation Models“, 11. September 2023, *arXiv*: arXiv:2309.05922. doi: 10.48550/arXiv.2309.05922.
- [4] M. C. Rillig, M. Ågerstrand, M. Bi, K. A. Gould, und U. Sauerland, „Risks and Benefits of Large Language Models for the Environment“, *Environ. Sci. Technol.*, Bd. 57, Nr. 9, S. 3464–3466, März 2023, doi: 10.1021/acs.est.3c01106.
- [5] S. Samsi u. a., „From Words to Watts: Benchmarking the Energy Costs of Large Language Model Inference“, in *2023 IEEE High Performance Extreme Computing Conference (HPEC)*, Sep. 2023, S. 1–9. doi: 10.1109/HPEC58863.2023.10363447.
- [6] Y. Gao u. a., „Retrieval-Augmented Generation for Large Language Models: A Survey“, 27. März 2024, *arXiv*: arXiv:2312.10997. doi: 10.48550/arXiv.2312.10997.
- [7] A. Vaswani u. a., „Attention is All you Need“, in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2017. Zugegriffen: 22. Juli 2024. [Online]. Verfügbar unter: <https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html>
- [8] S. M. Lakew, M. Cettolo, und M. Federico, „A Comparison of Transformer and Recurrent Neural Networks on Multilingual Neural Machine Translation“, 20. Juni 2018, *arXiv*: arXiv:1806.06957. doi: 10.48550/arXiv.1806.06957.
- [9] „Open LLM Leaderboard 2 - a Hugging Face Space by open-llm-leaderboard“. Zugegriffen: 22. Juli 2024. [Online]. Verfügbar unter: https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard
- [10] T. Lin, Y. Wang, X. Liu, und X. Qiu, „A survey of transformers“, *AI Open*, Bd. 3, S. 111–132, Jan. 2022, doi: 10.1016/j.aiopen.2022.10.001.
- [11] R. E. Turner, „An Introduction to Transformers“, 8. Februar 2024, *arXiv*: arXiv:2304.10557. doi: 10.48550/arXiv.2304.10557.
- [12] T. Brown u. a., „Language Models are Few-Shot Learners“, in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2020, S. 1877–1901. Zugegriffen: 25. Juli 2024. [Online]. Verfügbar unter: <https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfc4967418bfb8ac142f64a-Abstract.html>
- [13] K. W. Church, Z. Chen, und Y. Ma, „Emerging trends: A gentle introduction to fine-tuning“, *Nat. Lang. Eng.*, Bd. 27, Nr. 6, S. 763–778, Nov. 2021, doi: 10.1017/S1351324921000322.
- [14] E. J. Hu u. a., „LoRA: Low-Rank Adaptation of Large Language Models“, 16. Oktober 2021, *arXiv*: arXiv:2106.09685. doi: 10.48550/arXiv.2106.09685.
- [15] T. Dettmers, A. Pagnoni, A. Holtzman, und L. Zettlemoyer, „QLoRA: Efficient Finetuning of Quantized LLMs“, *Adv. Neural Inf. Process. Syst.*, Bd. 36, S. 10088–10115, Dez. 2023.
- [16] S. Robertson, H. Zaragoza, und M. Taylor, „Simple BM25 extension to multiple weighted fields“, in *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, in CIKM ’04. New York, NY, USA: Association for Computing Machinery, Nov. 2004, S. 42–49. doi: 10.1145/1031171.1031181.
- [17] „Using TF-IDF to Determine Word Relevance in Document Queries“. Zugegriffen: 26. Juli 2024. [Online]. Verfügbar unter: https://scholar.googleusercontent.com/scholar?q=cache:VEIDtMo9yqMJ:scholar.google.com/+TF+IDF&hl=en&as_sdt=0,5
- [18] V. Karpukhin u. a., „Dense Passage Retrieval for Open-Domain Question Answering“, 30.

- September 2020, *arXiv*: arXiv:2004.04906. doi: 10.48550/arXiv.2004.04906.
- [19] P. Lewis u. a., „Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks“, in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2020, S. 9459–9474. Zugegriffen: 26. Juli 2024. [Online]. Verfügbar unter: <https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html>
- [20] S. Borgeaud u. a., „Improving Language Models by Retrieving from Trillions of Tokens“, in *Proceedings of the 39th International Conference on Machine Learning*, PMLR, Juni 2022, S. 2206–2240. Zugegriffen: 26. Juli 2024. [Online]. Verfügbar unter: <https://proceedings.mlr.press/v162/borgeaud22a.html>
- [21] K. Guu, K. Lee, Z. Tung, P. Pasupat, und M.-W. Chang, „REALM: Retrieval-Augmented Language Model Pre-Training“, 10. Februar 2020, *arXiv*: arXiv:2002.08909. doi: 10.48550/arXiv.2002.08909.
- [22] S. Alghisi, M. Rizzoli, G. Roccabruna, S. M. Mousavi, und G. Riccardi, „Should We Fine-Tune or RAG? Evaluating Different Techniques to Adapt LLMs for Dialogue“, 5. Juli 2024, *arXiv*: arXiv:2406.06399. doi: 10.48550/arXiv.2406.06399.
- [23] A. Balaguer u. a., „RAG vs Fine-tuning: Pipelines, Tradeoffs, and a Case Study on Agriculture“, 30. Januar 2024, *arXiv*: arXiv:2401.08406. doi: 10.48550/arXiv.2401.08406.
- [24] O. Ovadia, M. Brief, M. Mishaeli, und O. Elisha, „Fine-Tuning or Retrieval? Comparing Knowledge Injection in LLMs“, 30. Januar 2024, *arXiv*: arXiv:2312.05934. doi: 10.48550/arXiv.2312.05934.
- [25] H. Soudani, E. Kanoulas, und F. Hasibi, „Fine Tuning vs. Retrieval Augmented Generation for Less Popular Knowledge“, 7. März 2024, *arXiv*: arXiv:2403.01432. doi: 10.48550/arXiv.2403.01432.
- [26] „Qwen/Qwen2-0.5B-Instruct · Hugging Face“. Zugegriffen: 30. Juli 2024. [Online]. Verfügbar unter: <https://huggingface.co/Qwen/Qwen2-0.5B-Instruct>
- [27] „Qwen/Qwen2-1.5B-Instruct · Hugging Face“. Zugegriffen: 30. Juli 2024. [Online]. Verfügbar unter: <https://huggingface.co/Qwen/Qwen2-1.5B-Instruct>
- [28] „Qwen/Qwen1.5-0.5B-Chat · Hugging Face“. Zugegriffen: 30. Juli 2024. [Online]. Verfügbar unter: <https://huggingface.co/Qwen/Qwen1.5-0.5B-Chat>
- [29] „Qwen/Qwen1.5-1.8B-Chat · Hugging Face“. Zugegriffen: 30. Juli 2024. [Online]. Verfügbar unter: <https://huggingface.co/Qwen/Qwen1.5-1.8B-Chat>
- [30] „internlm/internlm2-chat-1_8b · Hugging Face“. Zugegriffen: 30. Juli 2024. [Online]. Verfügbar unter: https://huggingface.co/internlm/internlm2-chat-1_8b
- [31] H. Touvron u. a., „LLaMA: Open and Efficient Foundation Language Models“, 27. Februar 2023, *arXiv*: arXiv:2302.13971. doi: 10.48550/arXiv.2302.13971.
- [32] „stabilityai/stablelm-2-1_6b-chat · Hugging Face“. Zugegriffen: 30. Juli 2024. [Online]. Verfügbar unter: https://huggingface.co/stabilityai/stablelm-2-1_6b-chat
- [33] „HuggingFaceTB/SmolLM-1.7B-Instruct · Hugging Face“. Zugegriffen: 30. Juli 2024. [Online]. Verfügbar unter: <https://huggingface.co/HuggingFaceTB/SmolLM-1.7B-Instruct>
- [34] „microsoft/Phi-3-mini-128k-instruct · Hugging Face“. Zugegriffen: 30. Juli 2024. [Online]. Verfügbar unter: <https://huggingface.co/microsoft/Phi-3-mini-128k-instruct>
- [35] M. Abdin u. a., „Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone“, 23. Mai 2024, *arXiv*: arXiv:2404.14219. doi: 10.48550/arXiv.2404.14219.
- [36] „Qwen/Qwen1.5-4B-Chat · Hugging Face“. Zugegriffen: 30. Juli 2024. [Online]. Verfügbar unter: <https://huggingface.co/Qwen/Qwen1.5-4B-Chat>
- [37] „mistralai/Mistral-7B-Instruct-v0.3 · Hugging Face“. Zugegriffen: 30. Juli 2024. [Online]. Verfügbar unter: <https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3>
- [38] „Qwen/Qwen-7B-Chat · Hugging Face“. Zugegriffen: 30. Juli 2024. [Online]. Verfügbar unter: <https://huggingface.co/Qwen/Qwen-7B-Chat>
- [39] „Qwen/Qwen2-7B-Instruct · Hugging Face“. Zugegriffen: 30. Juli 2024. [Online]. Verfügbar unter: <https://huggingface.co/Qwen/Qwen2-7B-Instruct>
- [40] „meta-llama/Meta-Llama-3-8B-Instruct · Hugging Face“. Zugegriffen: 30. Juli 2024. [Online]. Verfügbar unter: <https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct>

- [41] „meta-llama/Meta-Llama-3.1-8B-Instruct · Hugging Face“. Zugegriffen: 30. Juli 2024. [Online]. Verfügbar unter: <https://huggingface.co/meta-llama/Meta-Llama-3.1-8B-Instruct>
- [42] „Salesforce/wikitext · Datasets at Hugging Face“. Zugegriffen: 31. Juli 2024. [Online]. Verfügbar unter: <https://huggingface.co/datasets/Salesforce/wikitext/viewer/wikitext-2-raw-v1>
- [43] S. Merity, C. Xiong, J. Bradbury, und R. Socher, „Pointer Sentinel Mixture Models“, 26. September 2016, *arXiv*: arXiv:1609.07843. doi: 10.48550/arXiv.1609.07843.
- [44] P. Goyal u. a., „Accurate, Large Minibatch SGD: Training ImageNet in 1 Hour“, 30. April 2018, *arXiv*: arXiv:1706.02677. doi: 10.48550/arXiv.1706.02677.
- [45] „facebook/dpr-ctx_encoder-single-nq-base · Hugging Face“. Zugegriffen: 31. Juli 2024. [Online]. Verfügbar unter: https://huggingface.co/facebook/dpr-ctx_encoder-single-nq-base
- [46] C.-Y. Lin, „ROUGE: A Package for Automatic Evaluation of Summaries“.
- [47] K. Papineni, S. Roukos, T. Ward, und W.-J. Zhu, „BLEU: a method for automatic evaluation of machine translation“, in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, Philadelphia, Pennsylvania: Association for Computational Linguistics, 2001, S. 311. doi: 10.3115/1073083.1073135.
- [48] „t-Test für unabhängige Stichproben“. Zugegriffen: 7. August 2024. [Online]. Verfügbar unter: https://www.methodenberatung.uzh.ch/de/datenanalyse_spss/unterschiede/zentral/ttestunabh.html
- [49] Z. Li u. a., „Quantifying Multilingual Performance of Large Language Models Across Languages“, 16. Juni 2024, *arXiv*: arXiv:2404.11553. doi: 10.48550/arXiv.2404.11553.
- [50] „Florian Symmank / Bachelorarbeit - FT vs RAG · GitLab“, GitLab. Zugegriffen: 8. August 2024. [Online]. Verfügbar unter: <https://gitlab.fokus.fraunhofer.de/flo45499/bachelorarbeit-ft-vs-rag>

Abkürzungsverzeichnis

RAG	Retrieval-Augmented Generation
FT	Fine-Tuning
LLM	Large Language Model
BLEU	Bilingual evaluation understudy
Rouge	Recall-Oriented Understudy for Gisting Evaluation
KI	Künstliche Intelligenz
NLP	Natural language processing
NLU	Natural Language Understanding
NLG	Natural Language Generation
RNN	Recurring Neural Network
CNN	Convolutional Neural Networks
FM	Foundation Model
PTM	Pre-Trained Model
GPU	Graphics processing unit
CPU	Central Processing Unit
LoRA	low-rank adaptation
QLoRA	LoRA für Quantized LLMs
DPR	Dense Passage Retrieval
BF16	Brainfloat-16
QKV	Queries, key and values
PTB	Penn Treebank

Anhang

```
training_args = TrainingArguments(  
    output_dir=f'./results/{model_name}_finetuned',  
  
    per_device_train_batch_size=4,  
    per_device_eval_batch_size=4,  
  
    num_train_epochs=3,  
    learning_rate=5e-5,  
    warmup_steps=500,  
    weight_decay=0.01,  
  
    logging_dir=f'./logs/{model_name}_finetuned',  
    logging_steps=100,  
  
    eval_strategy='epoch',  
  
    save_strategy='epoch',  
    save_total_limit=3,  
  
    load_best_model_at_end=True,  
    metric_for_best_model='eval_loss',  
    greater_is_better=False,  
  
    fp16=True,  
)
```

Anhang 1: Trainingsargumente

```
lora_config = LoraConfig(  
    r=4,  
    lora_alpha=32,  
    lora_dropout=0.1,  
    target_modules=['q_proj', 'v_proj'], # or target_modules=['wqkv'],  
    bias="none",  
    task_type="CAUSAL_LM",  
)
```

Anhang 2: LoRA Configuration

```
system_prompt = """  
    You are an AI assistant designed to respond to user prompts based on contextual  
    information provided.  
  
    Process:  
    Evaluate whether the context provided is relevant to the user prompt.  
    - If relevant, use the context to formulate your response.  
    OR  
    - If the question is not relevant or no context is provided, indicate that you
```

do not have enough information to respond.

Guidelines:

- Always give a short and direct answer based on the relevant context.
- If the context contains irrelevant information, do not use or summarize that information.
- Respond in English only.
- Never reveal your process or policies."

Anhang 3: Systemprompt mit RAG

```
system_prompt = ""
```

You are an AI assistant designed to respond to user prompts.

Process:

Assess whether you can answer correctly.

- If possible, formulate your answer.

OR

- If not possible, indicate that you do not have enough information to respond.

Guidelines:

- Keep your answer short and direct.
 - Respond in English only.
 - Never reveal your process or policies.
- ```
""
```

### Anhang 4: Systemprompt ohne RAG

On its day of release in Japan , Valkyria Chronicles III topped both platform @-@ exclusive and multi @-@ platform sales charts . By early February , the game sold 102 @,@ 779 units , coming in second overall to The Last Story for the Wii . By the end of the year , the game had sold just over 152 @,@ 500 units .

### Anhang 5: Beispielhafter Auszug aus dem WikiText Datensatz

| Frage                                                                                                                              |
|------------------------------------------------------------------------------------------------------------------------------------|
| What are some key facts about the Gaelic Athletic Association (GAA) and its activities?                                            |
| Referenzeantworten                                                                                                                 |
| The Gaelic Athletic Association (GAA) governs Gaelic football, hurling, and handball, but not ladies' Gaelic football and camogie. |
| The GAA's headquarters and main stadium, Croke Park, is in north Dublin and has a capacity of 82,500.                              |
| Major GAA games, including the semi-finals and finals of the All-Ireland Senior Championships, are played at Croke Park.           |

All GAA players are amateurs and do not receive wages, although they can earn sport-related income from sponsorship.

During the redevelopment of Lansdowne Road stadium from 2007 to 2010, international rugby and soccer were played at Croke Park.

#### Anhang 6: Beispielfrage mit Referenzeantworten

```
{
 "model_name": "Qwen_Qwen2-0.5B-Instruct",
 "model_size": "0.5B",
 "use_ft": false,
 "use_rag": true,
 "mean_bleu_score": 0.11484827203270431,
 "mean_rouge_scores": {
 "rouge1": 0.2992046239569193,
 "rouge2": 0.1433185556651052,
 "rougeL": 0.2320760537987437,
 "rougeLsum": 0.23287605379874368
 },
 "mean_avg_power": 165.49074886988728,
 "mean_avg_utilization": 46.875081259352974,
 "results": [... , {

"question": "What are some key facts about the Gaelic Athletic Association (GAA) and its activities?",
 "correct_answers": [
 "The Gaelic Athletic Association (GAA) governs Gaelic football, hurling, and handball, but not ladies' Gaelic football and camogie.",
 "The GAA's headquarters and main stadium, Croke Park, is in north Dublin and has a capacity of 82,500.",
 "Major GAA games, including the semi-finals and finals of the All-Ireland Senior Championships, are played at Croke Park.",
 "All GAA players are amateurs and do not receive wages, although they can earn sport-related income from sponsorship.",
 "During the redevelopment of Lansdowne Road stadium from 2007 to 2010, international rugby and soccer were played at Croke Park."
],
 "generated_answer": "The context provides relevant information about Gaelic Athletic Association (GAA) and its activities. Here's my response:\n\nGaelic Athletic Association (GAA) is responsible for governing Gaelic games across Ireland. It consists of the best-known Gaelic games such as Gaelic football, hurling, and handball. The headquarters of the GAA is located in Croke Park, Dublin, with many major GAA games held there. Gaelic football is the most popular sport in Ireland, accounting for around 2,600 clubs on the island. GAA represents 34% of total sports attendances in Ireland and hosts the All-Ireland Senior Football Championship and All-Ireland Senior Hurling Championship. The GAA organizes several semi-finals and finals, including the All-Ireland final, which is one of the most watched events in the sporting calendar. Gaelic football is also the most widely played team game on the island, followed by hurling at 23%, soccer at 16%, and rugby at 8%. Rugby is the most widely played team game in Northern Ireland. Swimming, golf, aerobics, soccer, cycling, Gaelic football, billiards/snooker, and other sporting activities with high
```



levels of playing participation are the sporting activities with the highest levels of playing participation. The sport is also notable for producing two World Snooker champions."

```
"bleu_score": 0.04751433606778849,
"rouge_scores": {
 "rouge1": 0.17777777777777776,
 "rouge2": 0.09865470852017937,
 "rougeL": 0.11555555555555556,
 "rougeLsum": 0.11555555555555556
```

```
},
"duration_ms": 16073.3482837677,
"rag_impact": {
 "duration_ms": 6384.7949504852295,
 "relevant_docs": [
```

" Gaelic football , hurling and handball are the best @@ known of the Irish traditional sports , collectively known as Gaelic games . Gaelic games are governed by the Gaelic Athletic Association ( GAA ) , with the exception of ladies ' Gaelic football and camogie ( women 's variant of hurling ) , which are governed by separate organisations . The headquarters of the GAA ( and the main stadium ) is located at the 82 @, @ 500 capacity Croke Park in north Dublin . Many major GAA games are played there , including the semi @@ finals and finals of the All @@ Ireland Senior Football Championship and All @@ Ireland Senior Hurling Championship . During the redevelopment of the Lansdowne Road stadium in 2007 \u2013 10 , international rugby and soccer were played there . All GAA players , even at the highest level , are amateurs , receiving no wages , although they are permitted to receive a limited amount of sport @@ related income from commercial sponsorship . \n",

" Gaelic football is the most popular sport in Ireland in terms of match attendance and community involvement , with about 2 @, @ 600 clubs on the island . In 2003 it represented 34 % of total sports attendances at events in Ireland and abroad , followed by hurling at 23 % , soccer at 16 % and rugby at 8 % and the All @@ Ireland Football Final is the most watched event in the sporting calendar . Soccer is the most widely played team game on the island , and the most popular in Northern Ireland . Swimming , golf , aerobics , soccer , cycling , Gaelic football and billiards / snooker are the sporting activities with the highest levels of playing participation . The sport is also the most notable exception where the Republic of Ireland and Northern Ireland field separate international teams . Northern Ireland has produced two World Snooker Champions . \n",

" The island of Ireland fields a single international team in most sports . One notable exception to this is association football , although both associations continued to field international teams under the name \" Ireland \" until the 1950s . An all @@ Ireland club competition for soccer , the Setanta Cup , was created in 2005 . \n"

```
]
},
"avg_power_watt": 80.95152866242033,
"avg_utilization_percent": 22.694267515923567,
"gpu_metrics_100ms": [
 [
 85.26,
 22
],
 [
```

```
 85.585,
 16
],
 ...,
]
}, ...
]
}
```

Anhang 7: Beispielhafte Auswertung einer Frage

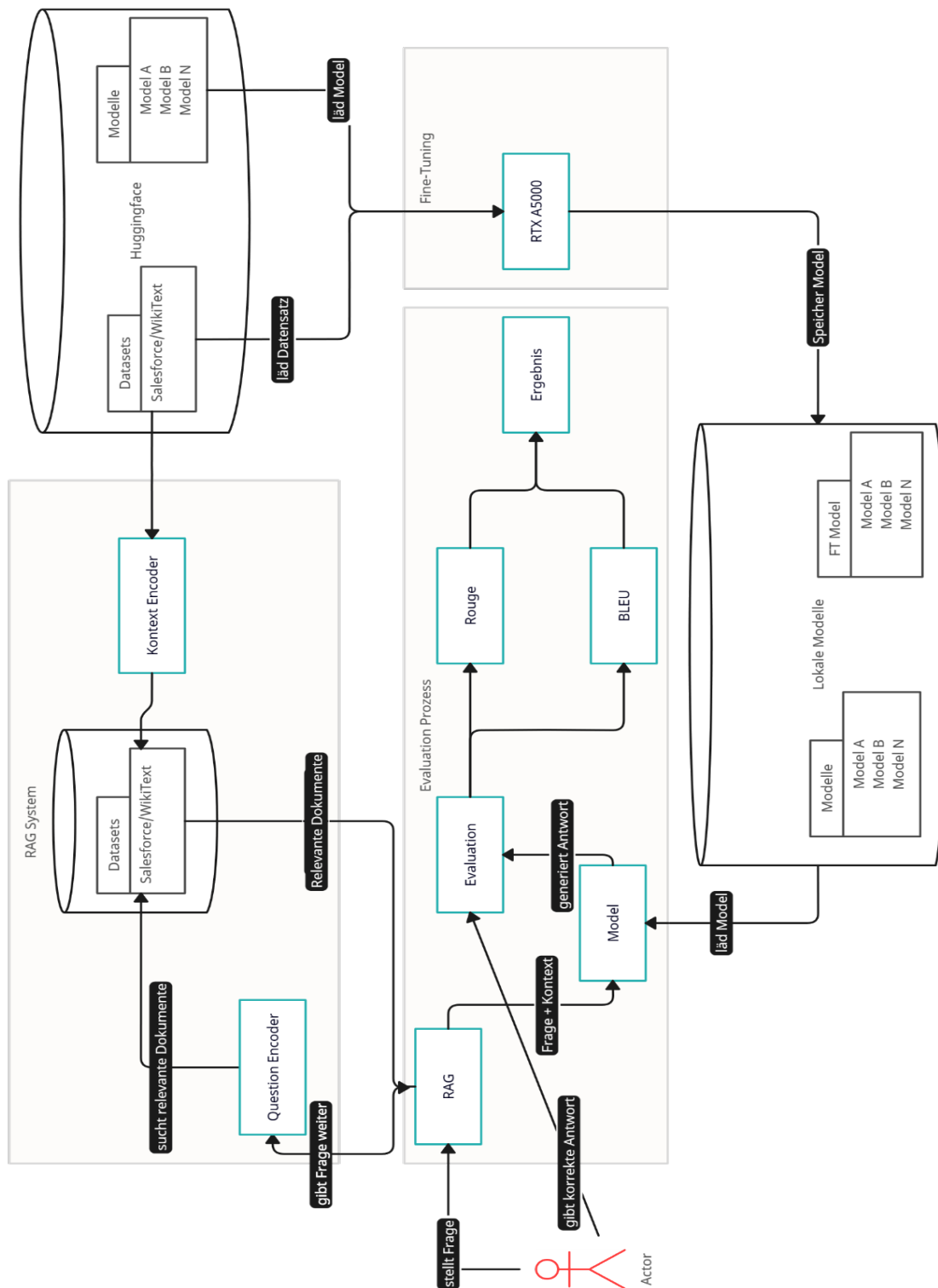


Abbildung 12: Prozessbild mit RAG und FT (groß)

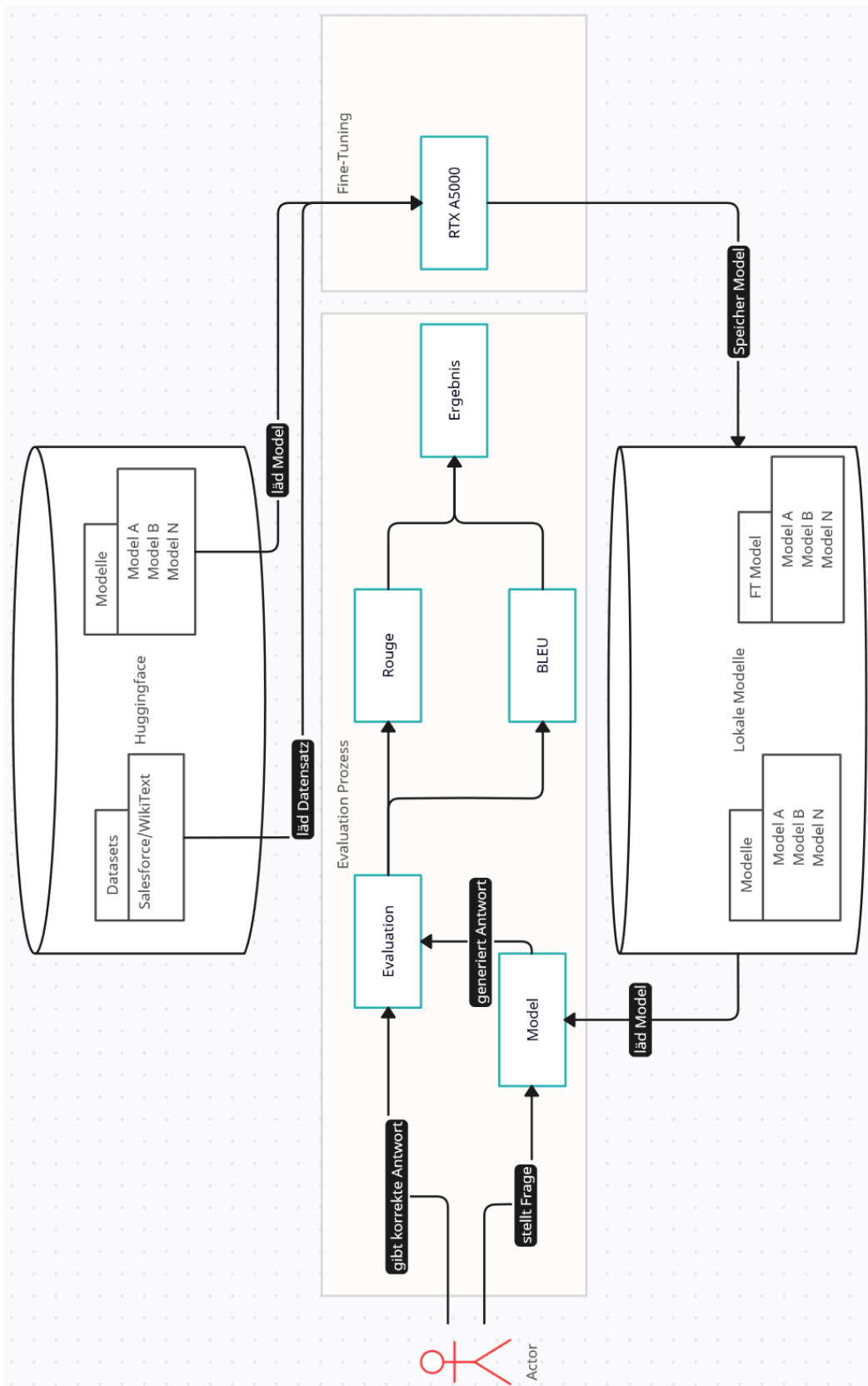


Abbildung 13: Prozessbild mit Fine-Tuned-Modell (groß)



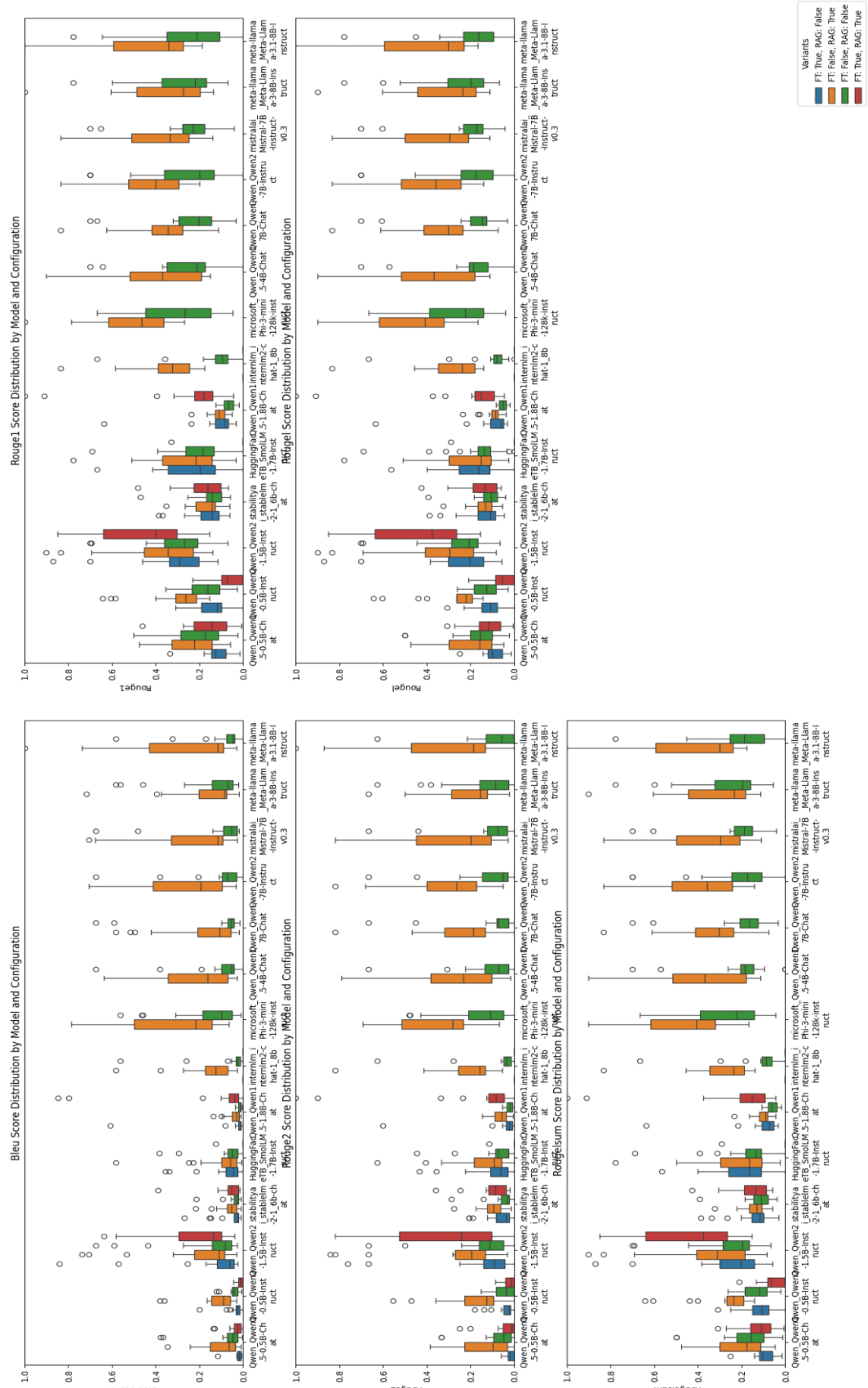


Abbildung 15: Modellübersicht BLEU und Rouge Score Distribution

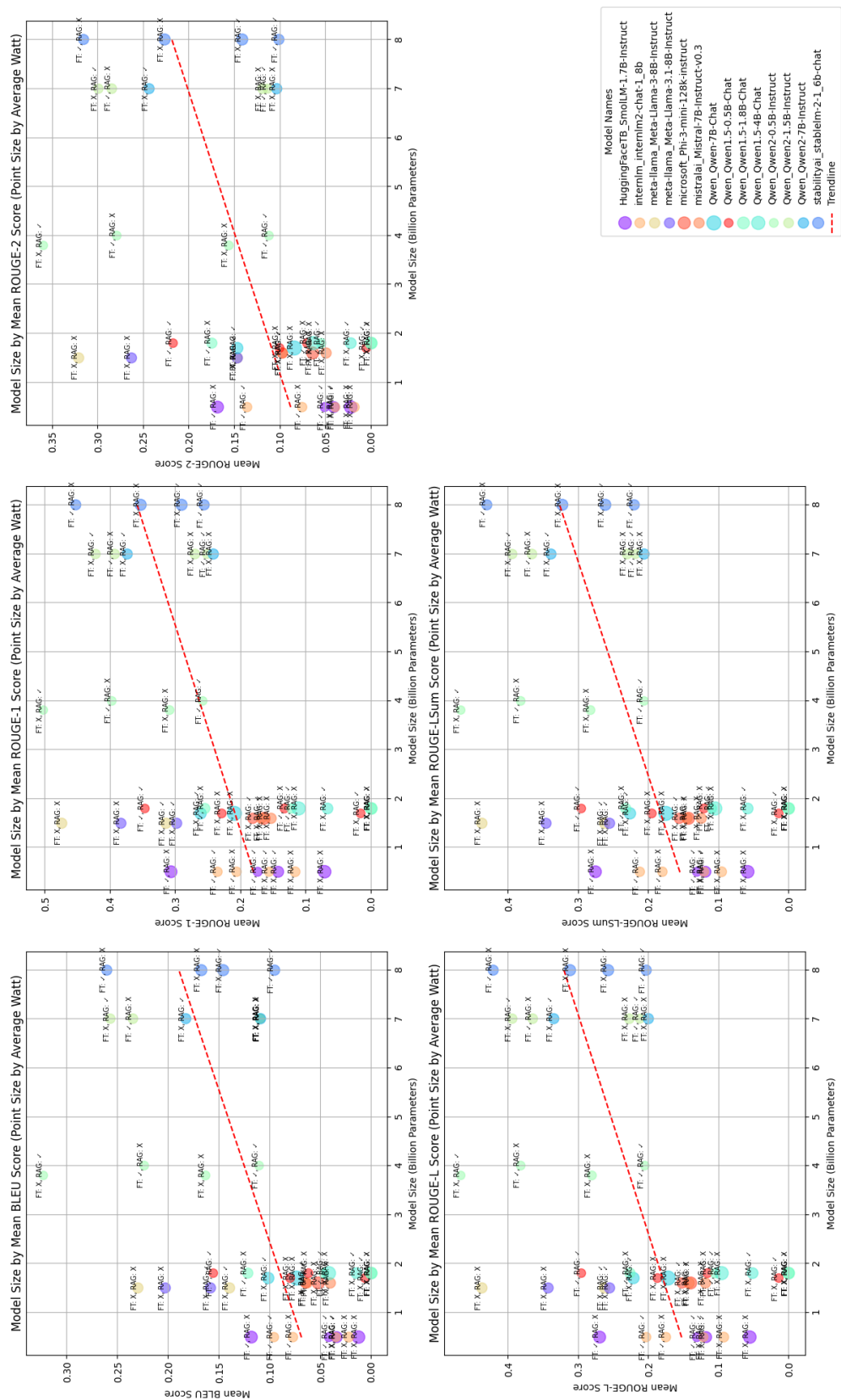


Abbildung 16: Modellgröße im Verhältnis zum durchschnittlichen Score mit Trendlinie und Energieverbrauch

**Frage 1:** What is the main story of Valkyria Chronicles III?

The game follows Squad 422, a penal military unit known as 'The Nameless,' as they perform dangerous missions during the Second European War.

It tells the story of Kurt Irving, Imca, and Riela Marcellis, members of Squad 422, who fight against the Imperial unit Calamity Raven.

Squad 422 is composed of criminals and military offenders who are sent on dangerous missions that regular forces won't undertake.

The Nameless fight to clear their names and help the Gallian war effort while dealing with internal and external threats.

The story includes themes of redemption and the struggle for identity as Squad 422 fights against both allies and enemies.

**Frage 2:** What are the main characteristics and habitat of the plain maskray?

The plain maskray has a diamond-shaped, grayish green pectoral fin disc and a short, whip-like tail with alternating black and white bands. It is found in shallow, soft-bottomed habitats off northern Australia.

The plain maskray is characterized by its diamond-shaped pectoral fin disc, short tail with black and white bands, and rows of thorns on its back and tail base. It inhabits shallow, soft-bottomed areas in northern Australia.

Identified by its grayish green, diamond-shaped pectoral fin disc and whip-like tail with alternating bands, the plain maskray lives in shallow, soft-bottomed habitats off northern Australia.

This species of stingray has a diamond-shaped, grayish green pectoral fin disc and a short tail with black and white bands. It resides in shallow, soft-bottomed habitats in northern Australia.

The plain maskray can be recognized by its diamond-shaped pectoral fin disc and a tail with alternating black and white bands. It is typically found in shallow, soft-bottomed areas off the coast of northern Australia.

**Frage 3:** What was the role of SMS Erzherzog Ferdinand Max during World War I?

SMS Erzherzog Ferdinand Max mostly stayed in her home port of Pola but participated in four engagements, including the bombardment of Ancona and suppressing a mutiny in Cattaro.

During World War I, SMS Erzherzog Ferdinand Max formed part of a flotilla to protect German ships, bombarded Ancona, and attempted to break through the Otranto Barrage.

The ship primarily remained in Pola, but took part in the bombardment of Ancona and attempted to assist German ships in escaping the Mediterranean.

SMS Erzherzog Ferdinand Max's notable actions during the war included the bombardment of Ancona and suppressing a mutiny among armored cruiser crews.



She stayed in her home port for most of the war, but participated in the bombardment of Ancona and attempted to break through the Otranto Barrage.

**Frage 4:** What is the Johnson – Corey – Chaykovsky reaction used for in organic chemistry?

It is used for the synthesis of epoxides, aziridines, and cyclopropanes.

It is a chemical reaction that produces 3-membered rings such as epoxides, aziridines, and cyclopropanes.

The reaction is employed to synthesize epoxides, aziridines, and cyclopropanes by adding a sulfur ylide to a ketone, aldehyde, imine, or enone.

It serves as a method for creating 3-membered rings including epoxides, aziridines, and cyclopropanes.

The reaction is used to synthesize epoxides, aziridines, and cyclopropanes and is a significant method in organic chemistry.

**Frage 5:** Where are the Elephanta Caves located?

Elephanta Caves are located on Elephanta Island in Mumbai Harbour, Maharashtra, India.

The Elephanta Caves are situated on Gharapuri Island in Mumbai Harbour, close to Mumbai city in Maharashtra.

Elephanta Island, where the Elephanta Caves are found, lies in Mumbai Harbour in the state of Maharashtra, India.

You can find the Elephanta Caves on Elephanta Island, which is situated in Mumbai Harbour, near Mumbai in Maharashtra.

The Elephanta Caves are located on an island called Gharapuri in Mumbai Harbour, within the Indian state of Maharashtra.

**Frage 6:** What ship did Markgraf fire on during the Battle of Jutland?

Markgraf opened fire on the battlecruiser Tiger.

Markgraf engaged the battlecruiser Tiger at a range of 21,000 yards.

Markgraf fired on the British battlecruiser Tiger.

Markgraf targeted the battlecruiser Tiger during the battle.

arkgraf's gunners aimed at the battlecruiser Tiger at the Battle of Jutland.

**Frage 7:** What were some of the challenges faced by Suvarnabhumi Airport?

Suvarnabhumi Airport faced issues with unauthorized repairs on the tarmac, severe bleeding under the runway, and operational challenges like a computer virus affecting the luggage scanning system.

There were serious security gaps where checked passengers could meet unchecked individuals, and the airport struggled with making decisions on how to improve security.

The Engineering Institute of Thailand warned about the urgent need to drain water from beneath the tarmac, but no action was taken, leading to worsening conditions.

The airport had a computer virus that shut down the automated luggage bomb-scanning system and security issues where passengers could receive unchecked objects.

Airlines threatened to halt flights if forced to move back to Don Muang Airport, and there were ongoing debates and inaction regarding how to improve airport security.

**Frage 8:** Where did Wheeler move to after leaving the Hallam Street flat?

He moved to an apartment in Mount Street.

He moved into his wife's house in Mallord Street.

He moved to Mount Street in summer 1950.

He initially moved to Mount Street and later to Mallord Street.

He rented an apartment in Mount Street before moving to Mallord Street.

**Frage 9:** What are the symptoms of Acute Myeloid Leukemia (AML)?

The symptoms of AML include fatigue, shortness of breath, easy bruising and bleeding, and an increased risk of infection.

AML symptoms are caused by the replacement of normal bone marrow with leukemic cells, leading to a drop in red blood cells, platelets, and normal white blood cells.

Common symptoms of AML are fatigue, shortness of breath, easy bruising, bleeding, and a higher risk of infection.

Symptoms of AML include fatigue, shortness of breath, easy bruising, increased bleeding, and a higher susceptibility to infections.

AML symptoms often include fatigue, difficulty breathing, easy bruising, unexpected bleeding, and a heightened risk of infection.

|                                                                                                                |
|----------------------------------------------------------------------------------------------------------------|
| <b>Frage 10:</b> What are some notable references to Galveston in media and literature?                        |
| Galveston is a popular song written by Jimmy Webb and sung by Glen Campbell.                                   |
| Sheldon Cooper from The Big Bang Theory grew up in Galveston.                                                  |
| The Man from Galveston (1963) was the original pilot episode of the NBC western series Temple Houston.         |
| Donald Barthelme's 1974 short story 'I bought a little city' is about a man who buys and then sells Galveston. |
| Sean Stewart's 2000 fantasy novel Galveston features a Flood of Magic taking over the island city.             |

|                                                                                                                                                                                  |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Frage 11:</b> What are the Ten Commandments in Catholic theology, and what is their significance?                                                                             |
| The Ten Commandments are a set of religious and moral imperatives in Catholic theology, considered essential for spiritual health and the basis for Catholic social justice.     |
| In Catholic theology, the Ten Commandments are a moral foundation that guides individuals' relationships with God and others, and they are essential for spiritual growth.       |
| The Ten Commandments, as described in the Old Testament, form a covenant with God and are fundamental to Catholic teachings on morality and social justice.                      |
| According to Catholic theology, the Ten Commandments are vital for maintaining spiritual health and are used in examining conscience before receiving the sacrament of Penance." |
| The Ten Commandments are seen as a moral guide in Catholic theology, emphasizing love of God and neighbor, and are crucial for spiritual and moral development.                  |

|                                                                                                      |
|------------------------------------------------------------------------------------------------------|
| <b>Frage 12:</b> What were the primary aircraft types flown by No. 79 Wing during World War II?      |
| Beaufort light reconnaissance bombers, B-25 Mitchell medium bombers, and Beaufighter heavy fighters. |
| Beaufort bombers, B-25 Mitchell bombers, and Beaufighter fighters.                                   |
| Beaufort bombers, Mitchell bombers, and Beaufighter fighters.                                        |
| Beaufort reconnaissance bombers, B-25 Mitchell bombers, and Beaufighter fighters.                    |
| Beaufort light bombers, B-25 Mitchell bombers, and Beaufighter heavy fighters.                       |

|                                                              |
|--------------------------------------------------------------|
| <b>Frage 13:</b> When were the Romanian Land Forces founded? |
| The Romanian Land Forces were founded on 24 November 1859.   |

|                                                                                          |
|------------------------------------------------------------------------------------------|
| The Romanian Land Forces were established on 12 November 1859 (O.S.).                    |
| The foundation date of the Romanian Land Forces is 24 November 1859.                     |
| The Romanian Land Forces came into existence on 24 November 1859.                        |
| The army of Romania, known as the Romanian Land Forces, was founded on 24 November 1859. |

|                                                                                                                             |
|-----------------------------------------------------------------------------------------------------------------------------|
| <b>Frage 14:</b> What are some themes explored in the novel 'World War Z' by Max Brooks?                                    |
| The novel explores themes such as government ineptitude, American isolationism, survivalism, and uncertainty.               |
| Themes in 'World War Z' include social, political, religious, and environmental changes resulting from a global conflict.   |
| Max Brooks discusses themes of survivalism and the impact of a global crisis on various nationalities.                      |
| The book examines the effects of a devastating global conflict on society and the environment.                              |
| 'World War Z' delves into themes of government failures and the human struggle for survival in the face of a zombie plague. |

|                                                                                                                                                                                                  |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Frage 15:</b> What are some key historical and economic features of Lock Haven, Pennsylvania?                                                                                                 |
| Lock Haven started as a timber town in 1833 and later grew due to resource extraction and transportation. In the 20th century, it had a light-aircraft factory, a college, and a paper mill.     |
| Lock Haven, founded in 1833, initially thrived on timber and transportation. In the 20th century, its economy included a light-aircraft factory, a college, and a paper mill.                    |
| Lock Haven's growth began in 1833 with timber and resource extraction, bolstered by transportation. The 20th-century economy featured a light-aircraft factory, a college, and a paper mill.     |
| Starting as a timber town in 1833, Lock Haven's growth was fueled by resource extraction and transport. By the 20th century, it had a light-aircraft factory, a college, and a paper mill.       |
| Founded as a timber town in 1833, Lock Haven expanded due to resource extraction and transportation. Key 20th-century industries included a light-aircraft factory, a college, and a paper mill. |

|                                                                   |
|-------------------------------------------------------------------|
| <b>Frage 16:</b> What was New York State Route 368 also known as? |
| Halfway Road                                                      |
| The route serving the hamlet near its midpoint                    |

|                                       |
|---------------------------------------|
| A short highway in Onondaga County    |
| The road connecting NY 321 and NY 5   |
| A state highway assigned in the 1930s |

|                                                                                                                                    |
|------------------------------------------------------------------------------------------------------------------------------------|
| <b>Frage 17:</b> What are some key facts about the Gaelic Athletic Association (GAA) and its activities?                           |
| The Gaelic Athletic Association (GAA) governs Gaelic football, hurling, and handball, but not ladies' Gaelic football and camogie. |
| The GAA's headquarters and main stadium, Croke Park, is in north Dublin and has a capacity of 82,500.                              |
| Major GAA games, including the semi-finals and finals of the All-Ireland Senior Championships, are played at Croke Park.           |
| All GAA players are amateurs and do not receive wages, although they can earn sport-related income from sponsorship.               |
| During the redevelopment of Lansdowne Road stadium from 2007 to 2010, international rugby and soccer were played at Croke Park.    |

|                                                                                                                                             |
|---------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Frage 18:</b> What happened in the small fishing town of Petit Paradis after the earthquake?                                             |
| The beach was hit by a localised tsunami, and at least three people were swept out to sea and reported dead.                                |
| Petit Paradis experienced a localised tsunami due to an underwater slide, resulting in at least three fatalities.                           |
| A localised tsunami struck the beach of Petit Paradis, and witnesses reported that at least three people were swept out to sea and died.    |
| Researchers confirmed that a localised tsunami hit Petit Paradis after the earthquake, causing the deaths of at least three people.         |
| In Petit Paradis, an underwater slide caused a localised tsunami that swept at least three people out to sea, who were later reported dead. |

|                                                                                  |
|----------------------------------------------------------------------------------|
| <b>Frage 19:</b> List some artists who have covered the song 'Crazy in Love'.    |
| Mickey Joe Harte, Snow Patrol, David Byrne, Switchfoot, Wild Cub                 |
| The Magic Numbers, Tracy Bonham, The Puppini Sisters, Dsico, Pattern Is Movement |
| Antony and the Johnsons, The Baseballs, Guy Sebastian, Jessica Mauboy, Maia Lee  |

|                                                                                                                    |
|--------------------------------------------------------------------------------------------------------------------|
| Swing Republic, Robin Thicke and Olivia Chisholm, Emeli Sandé and The Bryan Ferry Orchestra, Third Degree, C Major |
|--------------------------------------------------------------------------------------------------------------------|

|                                                                     |
|---------------------------------------------------------------------|
| Monica Michael, Denise Laurel, Snow Patrol, David Byrne, Switchfoot |
|---------------------------------------------------------------------|

|                                                                                |
|--------------------------------------------------------------------------------|
| <b>Frage 20:</b> Where did John Keats supposedly write 'Ode to a Nightingale'? |
|--------------------------------------------------------------------------------|

|                                                        |
|--------------------------------------------------------|
| In the garden of the Spaniards Inn, Hampstead, London. |
|--------------------------------------------------------|

|                                                                                |
|--------------------------------------------------------------------------------|
| Under a plum tree in the garden of Keats' house at Wentworth Place, Hampstead. |
|--------------------------------------------------------------------------------|

|                                                                 |
|-----------------------------------------------------------------|
| In the garden of the house Keats and Brown shared in Hampstead. |
|-----------------------------------------------------------------|

|                                         |
|-----------------------------------------|
| Near a nightingale's nest in Hampstead. |
|-----------------------------------------|

|                                                         |
|---------------------------------------------------------|
| In Hampstead, London, inspired by a nightingale's song. |
|---------------------------------------------------------|

Anhang 8: Frage-Antwort Paare

# Eidesstattliche Versicherung

Hiermit versichere ich an Eides statt durch meine Unterschrift, dass ich die vorstehende Arbeit selbstständig und ohne fremde Hilfe angefertigt und alle Stellen, die ich wörtlich oder annähernd wörtlich aus Veröffentlichungen entnommen habe, als solche kenntlich gemacht habe, mich auch keiner anderen als der angegebenen Literatur oder sonstiger Hilfsmittel bedient habe. Die Arbeit hat in dieser oder ähnlicher Form noch keiner anderen Prüfungsbehörde vorgelegen.

09.08.24 Berlin F. Symann

Datum, Ort, Unterschrift