

**Evaluierung der geographischen Fähigkeiten
ausgewählter Llama 3 Modelle:
Eine Untersuchung zur Genauigkeit der Koordinaten-
und Distanzermittlung von Städten**

BACHELORARBEIT
in teilweiser Erfüllung der Anforderungen für den Grad des
BACHELOR OF SCIENCE

Eingereicht von:

Justus Klameth

Matrikelnummer: 527 267

Betreuer:

Prof. Dr. Malte Schilling

Erstgutachter:

Prof. Dr. Malte Schilling

Zweitgutachter:

Prof. Dr. Xiaoyi Jiang

Münster, 5. Mai 2025

Zusammenfassung

Diese Arbeit untersucht die geographischen Fähigkeiten ausgewählter Llama 3 Modelle mit dem Ziel, deren Genauigkeit bei der Bestimmung von Stadtkoordinaten und Distanzen zwischen Städten zu analysieren. Dazu wurden mithilfe eines geeigneten Datensatzes verschiedene Anfrageformate und Auswertungsverfahren genutzt, um Antworten der Modelle zu sammeln. Diese wurden systematisch ausgewertet, um die Leistungsfähigkeit der Modelle zu ermitteln.

Die Ergebnisse zeigen, dass für die Stadtkoordinaten mit ausgewählten Modellen, Anfrage- und Auswertungsverfahren ein durchschnittlicher Fehler von unter 100 km erreicht werden kann, was im globalen Kontext als akzeptabel betrachtet werden könnte. Bei den Distanzen zwischen Städten liegen die durchschnittlichen Fehler bei über 600 km, was auch im globalen Maßstab als zu ungenau einzustufen ist. Besonders auffällig ist die hohe Standardabweichung in allen durchgeführten Experimenten. Dadurch sind die getesteten Verfahren für den praktischen Einsatz noch ungeeignet, da nicht nur ein möglichst geringer durchschnittlicher Fehler, sondern auch akkurate Einzelvorhersagen erforderlich sind.

Die Arbeit liefert damit einen Einblick in die geographischen Fähigkeiten ausgewählter Llama 3 Modelle und bildet eine Grundlage für zukünftige Untersuchungen und Vergleiche.

Inhaltsverzeichnis

1	Einleitung	1
2	Grundlagen	3
3	Methodik	7
3.1	Koordinaten	7
3.1.1	Datensatz	8
3.1.2	Anfragen	8
3.1.3	LLMs	8
3.1.4	Auswertung	9
3.2	Distanz	11
3.2.1	Datensatz	11
3.2.2	Anfragen	11
3.2.3	LLMs	11
3.2.4	Auswertung	12
4	Ergebnisse	13
4.1	Koordinaten	13
4.1.1	Quantitativer Vergleich der verschiedenen LLMs und Verfahren .	13
4.1.2	Qualitative Beispiele und Erörterung	18
4.2	Distanz	22
4.2.1	Quantitativer Vergleich der verschiedenen LLMs	22
4.2.2	Qualitative Beispiele und Erörterung	25
4.3	Schlüsselergebnisse beider Experimente	26
5	Diskussion und Fazit	27
5.1	Diskussion und Perspektiven für zukünftige Forschung	27
5.2	Zusammenfassung und Fazit	29
	Literaturverzeichnis	31
	Eigenständigkeitserklärung	37

1 | Einleitung

Seit den letzten Jahren spielen KI-Dienste eine wichtige Rolle in der Gesellschaft. Über 40 % der Befragten in der Generation Z (18 bis 27 Jahre) in Deutschland geben in einer Studie an, aktiv KI-Dienste zu nutzen (Horizont, 2023). Dabei sind weltweit die größten Anwendungsgebiete Contenterstellung, Kreatives Brainstorming und Austesten der Fähigkeiten (Capgemini, 2023).

Zu diesen KI-Diensten gehören auch Large Language Models (LLMs), die sich insbesondere durch ihre Leistungsfähigkeit im Bereich der natürlichen Sprachverarbeitung auszeichnen (Zhao et al., 2024). Auf Basis von *Attention is All you Need* (Vaswani et al., 2017) wurden in den zurückliegenden Jahren eine Vielzahl von LLMs wie z. B. GPT-4 (OpenAI et al., 2023), die Llama 3 Modelle (AI @ Meta, 2024) und Claude 3.5¹ publiziert. Diese Modelle lernen aus rein sprachlichen Zusammenhängen und es ist fraglich, inwieweit sich die Erkenntnisse generalisieren und übertragen lassen. Die Llama 3 Modelle weisen unter Anderem eine gute Leistung und vielversprechende Perspektiven für die Zukunft in den Bereichen Medizin (Li et al., 2023), Code-Generierung (Rozière et al., 2024) und Mathematik (Azerbayev et al., 2024) auf. Allerdings ist der Kenntnisstand über den Umfang des geografischen Wissens der Llama 3 Modelle derzeit noch gering.

Ein gutes Verständnis des geografischen Wissens und der Fähigkeiten der Llama 3 Modelle ist laut Roberts et al. (2023) relevant für:

1. **Sicherheit.** Mit zunehmender Leistungsfähigkeit von KI-Modellen gehen auch potenzielle Gefahren und Sicherheitsrisiken einher (Amodei et al., 2016). Daher ist es von Relevanz, die Fähigkeiten von Llama 3 zu kennen, um einen sicheren Einsatz zu gewährleisten.
2. **Fortschritt.** Ein gutes Verständnis ist für die stetige Verbesserung zukünftiger Modelle von wesentlicher Bedeutung.
3. **Anwendungsmöglichkeiten.** Für die gezielte Nutzung von Llama 3 ist ein umfassendes Verständnis des geografischen Wissens und der Fähigkeiten erforderlich,

¹<https://www.anthropic.com/>

wobei starke geografische Fähigkeiten den Einsatz z. B. in den Reise- und Navigationsbranchen ermöglichen würden.

Um mehr über das geografische Wissen von den Llama 3 Modellen herauszufinden, stellt sich diese Arbeit die Frage:

In welchem Umfang sind ausgewählte Llama 3 Modelle in der Lage, akkurate Koordinaten für Städte anzugeben und Distanzen zwischen Städten zu bestimmen?

Zur Beantwortung der aufgeworfenen Frage erfolgt zunächst eine quantitative Analyse und Gegenüberstellung verschiedener Llama 3 Modelle unter Verwendung eines Datensatzes. Darüber hinaus werden in einer qualitativen Untersuchung die am häufigsten auftretenden Fehlerquellen erörtert.

Dies erfolgt in mehreren Schritten. Zunächst werden in Kapitel 2 einige Grundlagen erörtert. Darauf aufbauend werden in Kapitel 3 die Methodik und die Experimente dieser Arbeit präsentiert. Die daraus resultierenden Ergebnisse werden in Kapitel 4 vorgestellt. Abschließend werden in Kapitel 5 die Ergebnisse diskutiert, ein Fazit gezogen und Perspektiven für zukünftige Arbeiten gegeben.

2 | Grundlagen

In diesem Kapitel werden einige grundlegende Konzepte vorgestellt, auf denen diese Arbeit aufbaut. Zunächst werden die wesentlichen Grundlagen zu LLMs erläutert, gefolgt von einem Überblick über das geographische Basiswissen. Abschließend werden die in dieser Arbeit genutzten Evaluationsmetriken sowie einige technische Grundlagen präsentiert.

Grundlagen zu Large Language Models (LLMs)

Mit *Large Language Models (LLMs)* sind meistens *Language Models* basierend auf Transformatoren gemeint, die Hunderte Milliarden (oder mehr) Parameter enthalten und mithilfe von sehr vielen Textdaten trainiert werden. LLMs verfügen über ausgeprägte Fähigkeiten, natürliche Sprache zu verstehen und komplexe Aufgaben (durch Textgenerierung) zu lösen (Zhao et al., 2024). Weitere Informationen über die Entwicklung, Fähigkeiten und wichtige Techniken in Bezug auf LLMs liefert die Arbeit *A Survey of Large Language Models* von Zhao et al. (2024).

Meta hat mittlerweile drei Versionen an LLMs veröffentlicht: LLaMA (Touvron et al., 2023a), Llama 2 (Touvron et al., 2023b) und Llama 3 (AI @ Meta, 2024). Details zu den einzelnen Versionen und Modellen können den zugehörigen Quellen entnommen werden.

Von den Llama 3 Modellen gibt es verschiedene Ausführungen: Llama 3, Llama 3.1, Llama 3.2 und Llama 3.3¹. Von diesen Modellarten bieten die Llama 3.1 Modelle mit den Parametergrößen 8 Mrd., 70 Mrd. und 405 Mrd. das größte Spektrum. Das Llama 3.3 Modell gibt es nur mit 70 Mrd. Parametern und die Llama 3.2 Modelle mit 1 Mrd. und 3 Mrd. Parametern. Außerdem existieren Llama 3.3 Modelle, die auch Bilder als Eingabe verwenden können.

Anfragen an LLMs erfolgen meist in Form von strukturierten Texteingaben. Für die Llama 3 Modelle gibt es beispielsweise vier vordefinierte Rollen für die Textabschnitte: **system**

¹Details: <https://github.com/meta-llama/llama-models/blob/main/models/>

setzt den Kontext (z. B. Regeln und allgemeine Informationen), **user** ist die Eingabe des Nutzers, **ipython** wird bei der Nutzung externer Werkzeuge eingesetzt und **assistant** enthält die Antwort des Modells. Weitere Informationen zur Nutzung der Llama Modelle sind unter <https://www.llama.com/docs/> verfügbar.

Zudem existieren auch alternative APIs, wie unter anderem die OpenAI API², die beispielsweise von DeepInfra³ genutzt wird. DeepInfra ermöglicht durch diese Standardisierung eine einfache Interaktion mit einer Vielzahl verschiedener LLMs.

Geographisches Basiswissen

Da diese Arbeit sich mit den geographischen Fähigkeiten von LLMs beschäftigt, ist ein grundlegendes Verständnis von einigen geographischen Konzepten wichtig. Im Folgenden werden die geographischen Koordinaten, Ländercodes und Berechnungsverfahren für geographische Distanzen vorgestellt.

Die geographischen Koordinaten, die in dieser Arbeit genutzt werden, sind Kugelkoordinaten, die eine Position auf der Erdoberfläche beschreiben. Die Koordinaten bestehen aus Längen- und Breitengrad. Weitere Informationen zu geographischen Koordinaten, wie z. B. zu verschiedenen Darstellungsmöglichkeiten und was genau Längen- und Breitengrad bedeuten, sind unter https://de.wikipedia.org/wiki/Geographische_Koordinaten zu finden.

Um die Distanz zwischen zwei geographischen Koordinaten zu berechnen, gibt es unter Anderem die Möglichkeit, die kürzeste Distanz dieser Punkte auf einer Kugeloberfläche zu bestimmen (auch Orthodrom genannt⁴). Da die Erde keine perfekte Kugel ist, werden genauere Ergebnisse erzielt, wenn man statt einer Kugel das WGS84-Ellipsoid zugrunde legt. Zur Berechnung kann dabei der Algorithmus von Karney (2013) verwendet werden.

Manchmal wird anstatt eines Ländernamens der zugehörige Ländercode nach der ISO-3166-1-Kodierliste (Ländercodeliste)⁵ verwendet. Die Standardisierung hat den Vorteil, dass Probleme mit verschiedenen Bezeichnungen für ein Land in verschiedenen Sprachen sowie allgemeine Eindeutigkeitsprobleme behoben werden.

²<https://platform.openai.com/docs/overview>

³<https://deepinfra.com/>

⁴<https://de.wikipedia.org/wiki/Orthodrome>

⁵<https://de.wikipedia.org/wiki/ISO-3166-1-Kodierliste>

Evaluationsmetriken

Um die Genauigkeit und Streuung der Antworten zu beschreiben, werden in dieser Arbeit zwei grundlegende statistische Maße verwendet: das **arithmetische Mittel** (Durchschnitt) und die **Standardabweichung**.

Das arithmetische Mittel⁶ (auch Mittelwert oder Durchschnitt genannt) ist die Summe der gegebenen Werte geteilt durch die Anzahl der Werte und gibt damit an, wie hoch der typische Wert eines Datensatzes ist.

$$\bar{x}_{arithm} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{x_1 + x_2 + \dots + x_n}{n}$$

In den nachfolgenden Experimenten wird in der Regel das arithmetische Mittel von Daten ermittelt, die an manchen Stellen keine Zahlen enthalten, da zuvor Fehler aufgetreten sind. Diese Werte werden bei der Berechnung ignoriert⁷.

Die Standardabweichung⁸ ist die Quadratwurzel der Varianz und eins der wichtigsten Streuungsmaße der Stochastik. Sie beschreibt, wie stark die einzelnen Datenpunkte um das arithmetische Mittel streuen.

$$\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}_{arithm})^2}$$

Auch bei der Berechnung der Standardabweichung werden häufig Daten mit Lücken betrachtet, die dabei ignoriert werden⁹.

Technische Grundlagen

In dieser Arbeit werden das JSON-Format sowie reguläre Ausdrücke zur Datenextraktion aus Texten verwendet.

Das JSON-Format wird hauptsächlich zur Speicherung und Übertragung strukturierter Daten eingesetzt. Ein einfaches Beispiel aus dem Kapitel 4 ist in der Abbildung 2.1 dargestellt. Dabei ist *distance* der Schlüssel, mit dem der Wert 6171.479892 abgerufen werden kann. Auf diese Weise können leicht einzelne Werte aus strukturierten Daten

⁶<https://de.wikipedia.org/wiki/Mittelwert>

⁷Mithilfe von: <https://numpy.org/doc/2.0/reference/generated/numpy.nanmean.html>

⁸[https://de.wikipedia.org/wiki/Varianz_\(Stochastik\)](https://de.wikipedia.org/wiki/Varianz_(Stochastik))

⁹Mithilfe von: <https://numpy.org/doc/2.1/reference/generated/numpy.nanstd.html>

```
1 {  
2   "distance": 6171.479892  
3 }
```

Abbildung 2.1: Die Antwort des Llama-3.1-8B-Instruct Modells für die Anfrage 4.14.

extrahiert werden. Weitere Informationen zum JSON-Format findet man z. B. unter <https://de.wikipedia.org/wiki/JSON>.

Reguläre Ausdrücke (Abkürzung: Regex) sind eine Möglichkeit, ein festes Muster zu beschreiben, mit dem unter anderem in einem Text gesucht werden kann. Zum Beispiel stellt der reguläre Ausdruck **(ab)*** alle Wörter dar, die aus beliebig vielen Wiederholungen von **ab** bestehen. Mit deutlich komplexeren regulären Ausdrücken kann man verschiedene Koordinatenformate in unstrukturierten Texten suchen und die Koordinaten anschließend extrahieren. Weitere Informationen über die Semantik und den Nutzen von regulären Ausdrücken sind unter https://de.wikipedia.org/wiki/Regul%C3%A4rer_Ausdruck zu finden.

3 | Methodik

Zur Beantwortung der Fragestellung werden zwei Experimente durchgeführt. Im ersten Experiment werden verschiedene Llama 3 Modelle nach den Koordinaten bestimmter Städte befragt. Die auf diese Weise erhaltenen Antworten werden anschließend mit den tatsächlichen Werten verglichen. Im zweiten Experiment soll die Distanz zwischen zwei Städten bestimmt werden. Auch diese wird mit der tatsächlichen Distanz verglichen.

Beide Experimente folgen der in Abbildung 3.1 dargestellten Grundstruktur. Zunächst wird ein passender Datensatz erstellt und daraus entsprechende Anfragen generiert. Diese Anfragen werden an verschiedene LLMs gestellt und die Antworten anschließend für die Analyse aufbereitet, die in Kapitel 4 vorgestellt wird.

Die praktische Umsetzung der Methodik, die gesammelten Daten und die Auswertungen sind zu finden unter: <https://github.com/JustusKlameth/ba-dev>.

3.1 Koordinaten

In diesem Experiment wird der Fokus darauf gelegt zu ermitteln, *in welchem Umfang ausgewählte Llama 3 Modelle in der Lage sind, akkurate Koordinaten für Städte anzugeben*. Zu diesem Zweck werden die Llama 3 Modelle ähnlich zur Vorgehensweise von

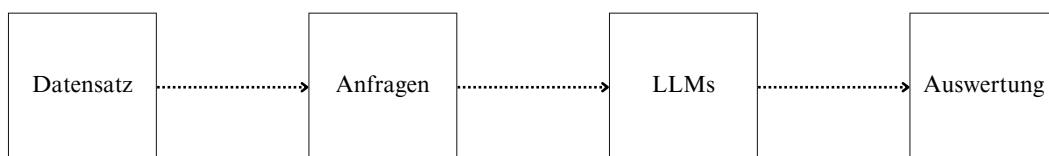


Abbildung 3.1: Grundstruktur der Experimente.

Bhandari et al. (2023) nach den Koordinaten von Städten befragt und die daraus resultierenden Antworten ausgewertet.

Darüber hinaus werden drei verschiedene Methoden getestet, um die angestrebten Informationen zu erlangen. Des Weiteren wird untersucht, welchen Einfluss es hat, zusätzlich zum Städtenamen auch das zugehörige Land zu übergeben.

3.1.1 Datensatz

Der Datensatz umfasst die global verteilten 3.527 Städte aus dem MaxMind Datensatz¹ mit mindestens 100.000 Einwohnern. Dieser Datensatz wurde ausgewählt, weil er alle für die Auswertung relevanten Informationen enthält (Städtename, Ländercode, Längen- und Breitengrad). Außerdem erlaubt dieser Datensatz einen Vergleich mit den Ergebnissen von Bhandari et al. (2023).

3.1.2 Anfragen

Die Generierung von Anfragen erfolgt unter Anwendung der Vorlagen 3.2 und 3.3, wobei für `{location}` der Städtename eingesetzt wird. Darüber hinaus besteht die Option, neben dem Städtenamen auch das zugehörige Land zu verwenden.

Für jede Stadt in dem Datensatz werden vier Anfragen erstellt, also für jede Vorlage einmal mit und einmal ohne die Länderinformation.

Die Vorlage 3.3 wurde ausgewählt, da Bhandari et al. (2023) auf diese Weise bereits vielversprechende Ergebnisse mit älteren Llama Modellen erzielt haben. Zusätzlich wurde die Vorlage 3.2 als Kontrast entwickelt, da sie weniger allgemeine Vorgaben enthält, dafür aber ein festes Antwortformat vorschreibt und damit einen neuen Ansatz verfolgt.

3.1.3 LLMs

Mit diesen Anfragen werden das Llama-3.3-70B-Instruct Modell und die Llama-3.1-Instruct Modelle in den Größen 8B, 70B und 405B getestet. Dafür wird DeepInfra² genutzt.

Die Llama 3.1 Modelle wurden ausgewählt, da sie die größte Spanne an Parametergrößen innerhalb einer Version liefern und somit einen umfassenden Vergleich ermöglichen.

¹<https://www.kaggle.com/datasets/max-mind/world-cities-database>

²<https://deepinfra.com/>

```

1  [
2    {
3      "role": "system",
4      "content": "Answer in json format only: {\\"latitude\\": < e.g.
                    \\"14.832185\\" >, \\"longitude\\": < e.g. \\"4.212666\\" >}. No
                    explanation, just the answer!"
5    },
6    {
7      "role": "user",
8      "content": "What are the latitude and longitude of {location}?"
9    }
10 ]

```

Abbildung 3.2: Diese Vorlage schreibt das *json*-Format für die Antwort vor. Dafür wird nicht explizit (wie bei 3.3) darauf hingewiesen, keine falschen Informationen zu verbreiten.

```

1  [
2    {
3      "role": "system",
4      "content": "You are a helpful, respectful and honest assistant. Always
                    answer as helpfully as possible, while being safe. Your answers
                    should not include any harmful, unethical, racist, sexist, toxic,
                    dangerous, or illegal content. Please ensure that your responses are
                    socially unbiased and positive in nature. If a question does not make
                    any sense, or is not factually coherent, explain why instead of
                    answering something not correct. If you don't know the answer to a
                    question, please don't share false information."
5    },
6    {
7      "role": "user",
8      "content": "What are the latitude and longitude of {location}?"
9    }
10 ]

```

Abbildung 3.3: Vorlage von Bhandari et al. (2023). Diese Vorlage nutzt eine offene Formulierung und schreibt kein Antwortformat vor.

Das Llama 3.3 Modell wurde zusätzlich betrachtet, um festzustellen, ob es bei gleicher Parameteranzahl eine Verbesserung des geographischen Wissens gibt.

3.1.4 Auswertung

Um die Qualität der Antworten messen zu können, wird als Fehler die geodätische Distanz zwischen den tatsächlichen Koordinaten der Stadt und den Koordinaten der Antwort des LLMs verwendet. Dafür wird der Algorithmus von Karney (2013) mit dem WGS-84-Ellipsoiden verwendet³.

³Implementierung: <https://geopy.readthedocs.io/en/stable/#module-geopy.distance>

```
1 Extract any explicit coordinates from the messages you receive.
2
3 - **Only** extract coordinates if they are explicitly mentioned in the
  message (e.g., numerical latitude and longitude).
4 - **Do not** infer or look up coordinates based on place names, city names,
  or any other information.
5 - **If multiple pairs of coordinates are present, extract only the first pair
  mentioned.**
6 - If there are no explicit coordinates in the message, reply "nan".
7
8 Answer in JSON format only: {"latitude": <e.g., "14.832185">, "longitude": <e
  .g., "4.212666">}. No explanation, just the answer!
```

Abbildung 3.4: System Prompt für das *llm*-Verfahren.

Zur Extraktion der Koordinaten aus der Antwort des LLMs werden drei verschiedene Verfahren genutzt.

Bei der Vorlage 3.2 wird das *json*-Format für die Antwort vorgeschrieben. Demzufolge können die Koordinaten extrahiert werden, indem die Antwort als *json*-Objekt interpretiert wird. Falls dies nicht möglich ist, wird die Antwort als fehlerhaft eingestuft. Diese Art der Auswertung wird im Folgenden als *json* bezeichnet.

Bei der Vorlage 3.3 wird kein Antwortformat vorgeschrieben. Daher sind die Antworten häufig Fließtexte. Um aus diesen Texten die Koordinaten zu extrahieren, werden zwei verschiedene Verfahren angewendet. Das erste Verfahren stammt von Bhandari et al. (2023) und wurde unverändert übernommen, um eine Vergleichbarkeit zu ermöglichen⁴. Dieses Verfahren basiert auf regulären Ausdrücken und wird daher im Folgenden als *regex* bezeichnet.

Das zweite Verfahren für Fließtexte liefert einen alternativen Ansatz zur Extraktion der Koordinaten und wurde speziell mit dem Ziel entwickelt, alle möglichen Darstellungen von Koordinaten zu erkennen. Um dies zu erreichen, wurde das Llama-3.1-8B-Instruct-Turbo Modell mit dem in Abbildung 3.4 dargestellten System Prompt genutzt und der Fließtext als Nutzereingabe verwendet. Die dadurch entstehende Antwort wird mit dem *json*-Verfahren weiterverarbeitet, da das System Prompt das *json*-Format vorschreibt. Dieses Verfahren wird im Folgenden als *llm* bezeichnet.

Des Weiteren ist darauf zu achten, dass für die *regex*- und *llm*-Verfahren stets die Vorlage 3.3 und für das *json*-Verfahren stets die Vorlage 3.2 verwendet wurde.

⁴https://github.com/prabin525/spatial-llm/blob/main/coor-prediction/calculate_errs.py

3.2 Distanz

In diesem Experiment wird der Frage nachgegangen, *in welchem Umfang ausgewählte Llama 3 Modelle in der Lage sind, Distanzen zwischen Städten zu bestimmen*. Zu diesem Zweck werden die Llama 3 Modelle anhand der im Datensatz enthaltenen Stadt-Paare befragt und die erhaltenen Antworten ausgewertet.

Darüber hinaus wird untersucht, welchen Einfluss die Hinzunahme des Landes auf die Genauigkeit hat.

3.2.1 Datensatz

Um einen geeigneten Datensatz zu erstellen, werden aus dem Datensatz 3.1.1 zufällige Paare gebildet. Dadurch entstehen 1.763 Paare von Städten mit mindestens 100.000 Einwohnern.

3.2.2 Anfragen

Die Anfragen werden durch die Anwendung der Vorlage 3.5 erstellt. Dafür werden für `{location_1}` und `{location_2}` die jeweiligen Städtenamen eingesetzt. Außerdem besteht die Möglichkeit, neben den Städtenamen auch die zugehörigen Länder einzusetzen.

Auf diese Weise werden für jedes Stadt-Paar im Datensatz zwei Anfragen erstellt, wobei die erste Anfrage lediglich die Städtenamen und die zweite die Städtenamen sowie die Länder beinhaltet.

Für dieses Experiment wird eine Vorlage zur Erstellung der Anfragen verwendet, die - ähnlich wie Vorlage 3.2 aus dem vorigen Experiment - das *json*-Format vorschreibt. Diese Entscheidung wurde getroffen, da das *json*-Verfahren praktisch immer auswertbare Antworten produziert und keine zusätzliche intransparente Komponente wie das *llm*-Verfahren einbringt (vgl. Kapitel 4 und 5).

3.2.3 LLMs

Wie beim vorigen Experiment, werden mit diesen Anfragen das Llama-3.3-70B-Instruct Modell und die Llama-3.1-Instruct Modelle in den Größen 8B, 70B und 405B getestet. Dafür wird DeepInfra⁵ genutzt.

⁵<https://deepinfra.com/>

```
1  [  
2    {  
3      "role": "system",  
4      "content": "Answer in json format only: {\"distance\": < e.g.  
        \"1945.399479\" >}. The value must be in kilometers and no  
        explanation, just the answer!"  
5    },  
6    {  
7      "role": "user",  
8      "content": "What is the distance between {location_1} and {location_2}?"  
9    }  
10 ]
```

Abbildung 3.5: Diese Vorlage schreibt das *json*-Format für die Antwort vor. Außerdem wird die Einheit Kilometer vorgegeben.

3.2.4 Auswertung

Die Qualität der Antworten wird gemessen, indem der absolute Unterschied zwischen der tatsächlichen geodätischen Distanz (basierend auf den realen Koordinaten der Städte) und der vom LLM angegebenen Distanz als Fehler verwendet wird. Zur Berechnung der geodätischen Distanz wird der Algorithmus von Karney (2013) mit dem WGS-84-Ellipsoiden verwendet⁶.

Zur Extraktion der Distanzen aus den Antworten des LLMs wird genutzt, dass die verwendete Vorlage 3.5 das *json*-Format für die Antworten vorschreibt. Dementsprechend werden die Antworten als *json*-Objekt interpretiert, sofern dies möglich ist. Andernfalls werden die Antworten als fehlerhaft markiert.

⁶Implementierung: <https://geopy.readthedocs.io/en/stable/#module-geopy.distance>

4 | Ergebnisse

In diesem Kapitel werden die Resultate präsentiert, die zur Beantwortung der folgenden Fragestellung dienen:

In welchem Umfang sind ausgewählte Llama 3 Modelle in der Lage, akkurate Koordinaten für Städte anzugeben und Distanzen zwischen Städten zu bestimmen?

Zunächst werden die Resultate des Experiments *Koordinaten* (3.1) präsentiert. Zu diesem Zweck werden zu Beginn die Fehler der verschiedenen Modelle und der verschiedenen Vorlagen ohne die Länderinformation quantitativ miteinander verglichen. Im Anschluss daran wird erörtert, welchen Einfluss die Hinzunahme des Landes hat. Abschließend werden qualitative Beispiele vorgestellt und häufige Fehler und Probleme präsentiert.

Im weiteren Verlauf werden die Resultate des Experiments *Distanz* (3.2) dargestellt. Für dieses Experiment erfolgt ebenso zunächst eine quantitative Analyse der Fehler der verschiedenen Modelle und der verschiedenen Vorlagen ohne Länderinformation. Im Anschluss daran wird der Einfluss der Länderinformation untersucht. Schließlich werden qualitative Beispiele vorgestellt und häufige Fehler und Probleme illustriert.

4.1 Koordinaten

Im Folgenden werden die Ergebnisse des Experiments *Koordinaten* dargestellt, die aus der in Kapitel 3.1 beschriebenen Vorgehensweise resultieren.

4.1.1 Quantitativer Vergleich der verschiedenen LLMs und Verfahren

Bei den Auswertungen und Visualisierungen in diesem Kapitel wird häufig der durchschnittliche Fehler dargestellt. Damit ist der durchschnittliche Fehler aller Antworten gemeint, die die angegebenen Bedingungen erfüllen, d. h. z. B. alle Antworten des Llama-3.3-70B-Instruct Modells mit dem *json*-Verfahren und den Länderinformationen. Dabei werden die fehlerhaften Antworten der LLMs nicht berücksichtigt.

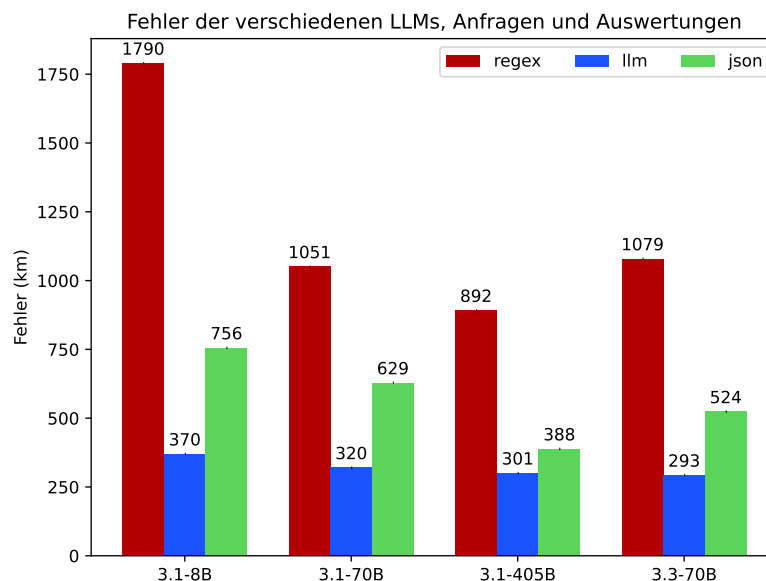


Abbildung 4.1: Der durchschnittliche Fehler von den verschiedenen LLMs, Anfragen und Auswertungsverfahren ohne die Länderinformation für das Experiment *Koordinaten* (3.1). Dabei werden die Fehler für die LLMs Llama-3.1-Instruct in den Größen 8B, 70B, 405B und Llama-3.3-70B-Instruct (links nach rechts) für jeweils die Auswertungsverfahren *regex* (rot), *llm* (blau) und *json* (grün) mit den zugehörigen Vorlagen visualisiert.

Ohne die Länderinformation

Der durchschnittliche Fehler der Anfragen ohne zusätzliche Länderinformationen zeigt, dass für alle betrachteten Modelle das *llm*-Verfahren am besten und das *regex*-Verfahren am schlechtesten ist, während das *json*-Verfahren zwischen den beiden anderen liegt. Dabei fällt auf, dass der Abstand zwischen dem *regex*- und dem *json*-Verfahren größer ist als zwischen dem *llm*- und dem *json*-Verfahren (vgl. Abbildung 4.1).

Außerdem zeigt die Abbildung 4.1, dass für alle betrachteten Verfahren die Genauigkeit bei größeren Modellen zunimmt. Die einzige Ausnahme stellt das Llama-3.3-70B-Instruct Modell mit dem *llm*-Verfahren dar, welches für dieses Verfahren die besten Ergebnisse liefert.

Der Anteil der verarbeitbaren Antworten für die Anfragen ohne Länderinformationen (vgl. Abbildung 4.2) zeigt, dass die Verfahren unabhängig vom Modell relativ konstante Fehlerraten aufweisen. Die Mittelwerte betragen für das *regex*-Verfahren $9,04 \pm 3,63$, für das *llm*-Verfahren $91,3 \pm 2,1$ und für das *json*-Verfahren $99,7 \pm 0,09$. Dabei fällt auf, dass das *json*-Verfahren nahezu fehlerfrei ist, während das *regex*-Verfahren im Durchschnitt über 90 Prozent der Anfragen so fehlerhaft beantwortet, dass es nicht möglich ist, Koordinaten zu extrahieren.

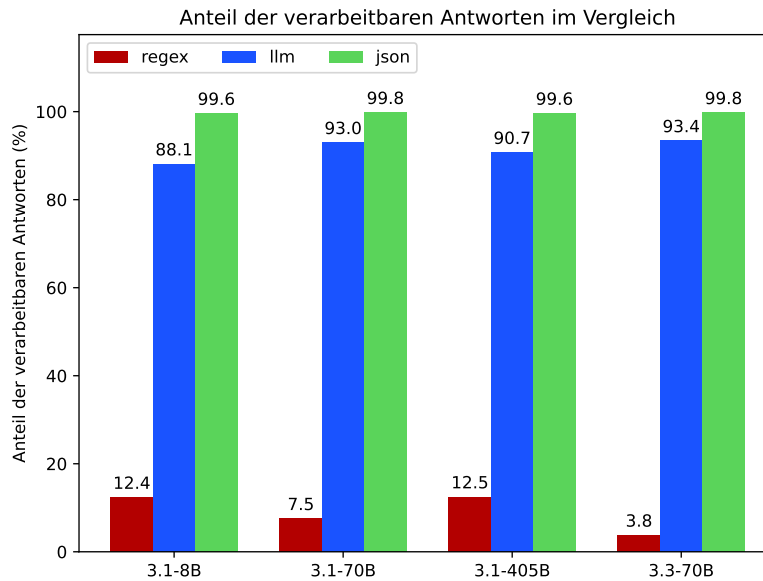


Abbildung 4.2: Der Anteil der korrekt verarbeitbaren Antworten der verschiedenen LLMs, Anfragen und Auswertungsverfahren für das Experiment *Koordinaten* (3.1) ohne die Länderinformationen. Dabei werden die Anteile der Antworten, die korrekt verarbeitet werden können, von den LLMs Llama-3.1-Instruct in den Größen 8B, 70B, 405B und Llama-3.3-70B-Instruct (links nach rechts) für jeweils die Auswertungsverfahren *regex* (rot), *llm* (blau) und *json* (grün) mit den zugehörigen Vorlagen visualisiert.

Die in den Abbildungen 4.1 und 4.2 dargestellten Ergebnisse weisen beide ein Muster auf. Bei jedem Modell hat das *regex*-Verfahren eindeutig den größten durchschnittlichen Fehler und bei weitem die meisten fehlerhaften Antworten. Im Kontrast dazu liegen das *json*- und das *llm*-Verfahren deutlich näher beieinander. Das *llm*-Verfahren liefert etwas bessere durchschnittliche Fehler, dafür das *json*-Verfahren aber nahezu keine fehlerhaften Antworten.

Allerdings entstehen bei diesem Verfahren teilweise große Fehler für einzelne Städte. Dies liegt unter anderem daran, dass Städtenamen nicht eindeutig sind. Das führt dazu, dass Koordinaten für andere Städte mit dem gleichen Namen zurückgegeben werden. Dies könnte der Grund dafür sein, dass z. B. bei dem Llama-3.1-405B-Instruct Modell mit dem *json*-Verfahren ohne die Länderinformation unter 1 % der Anfragen für über 35 % des Gesamtfehlers verantwortlich sind. In der Abbildung 4.10 ist z. B. dargestellt, dass bei diesen Anfragen die Stadt Münster (NRW, Deutschland) wahrscheinlich mit der südlichsten Provinz Munster in Irland¹ verwechselt wurde.

¹[https://de.wikipedia.org/wiki/Munster_\(Irland\)](https://de.wikipedia.org/wiki/Munster_(Irland))

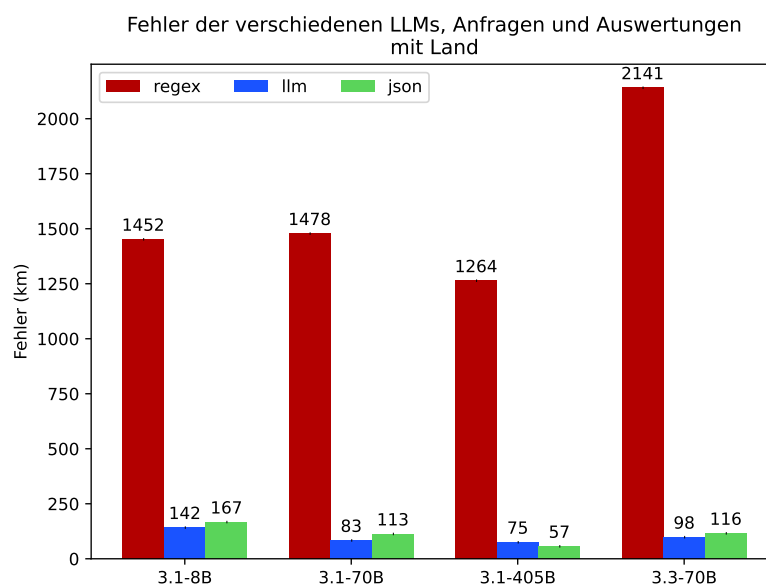


Abbildung 4.3: Der durchschnittliche Fehler von den verschiedenen LLMs, Anfragen und Auswertungsverfahren mit der Länderinformation für das Experiment *Koordinaten* (3.1). Dabei werden die Fehler für die LLMs Llama-3.1-Instruct in den Größen 8B, 70B, 405B und Llama-3.3-70B-Instruct (links nach rechts) für jeweils die Auswertungsverfahren *regex* (rot), *llm* (blau) und *json* (grün) mit den zugehörigen Vorlagen visualisiert.

Mit der Länderinformation

Daher wird im Folgenden vorgestellt, was passiert, wenn das Land zu den Anfragen hinzugefügt wird.

Der durchschnittliche Fehler der Anfragen mit der Länderinformation (vgl. Abbildung 4.3) zeigt ein ähnliches Muster wie der durchschnittliche Fehler der Anfragen ohne die Länderinformationen (vgl. Abbildung 4.1). Bei beiden Arten der Anfragen ist das *regex*-Verfahren am schlechtesten, wobei der Abstand bei den Anfragen mit der Länderinformation deutlich angestiegen ist. Außerdem schneidet bei beiden das *llm*-Verfahren am besten ab und das *json*-Verfahren liegt in der Mitte. Ein Ausreißer sind die Anfragen mit der Länderinformation für das Llama-3.1-405B-Instruct Modell. Hier liefert das *json*-Verfahren etwas bessere Ergebnisse als das *llm*-Verfahren.

Außerdem zeigt die Abbildung 4.3, dass bei den *json*- und *llm*-Verfahren die Genauigkeit der Antworten für größere Modelle zunimmt.

Der Anteil der verarbeitbaren Antworten verändert sich durch die Hinzunahme der Länderinformation kaum (vgl. Abbildungen 4.2 und 4.4). Dabei bleiben die Fehlerraten für die verschiedenen Verfahren unabhängig von der Modellgröße relativ konstant; das *regex*-Verfahren sehr fehleranfällig, das *llm*-Verfahren sehr gut und das *json*-Verfahren

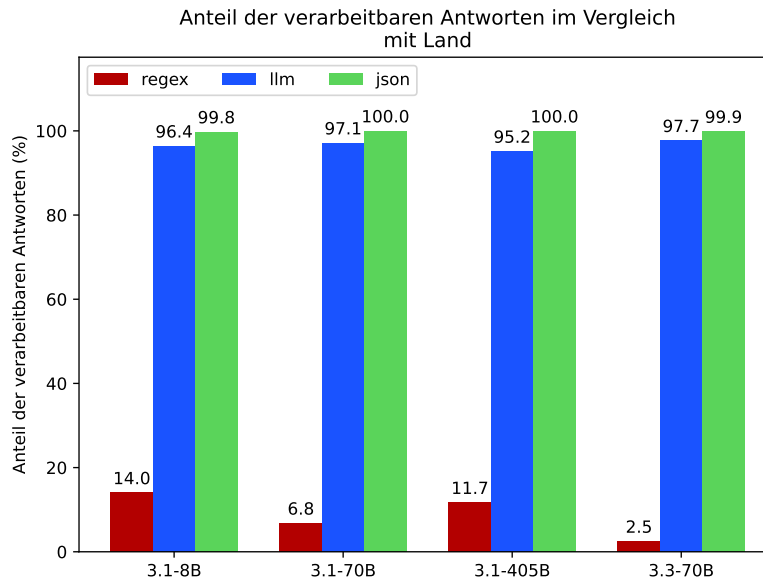


Abbildung 4.4: Der Anteil der korrekt verarbeitbaren Antworten der verschiedenen LLMs, Anfragen und Auswertungsverfahren für das Experiment *Koordinaten* (3.1) mit der Länderinformation. Dabei werden die Anteile der Antworten, die korrekt verarbeitet werden können, von den LLMs Llama-3.1-Instruct in den Größen 8B, 70B, 405B und Llama-3.3-70B-Instruct (links nach rechts) für jeweils die Auswertungsverfahren *regex* (rot), *llm* (blau) und *json* (grün) mit den zugehörigen Vorlagen visualisiert.

praktisch perfekt. Die Mittelwerte betragen für das *regex*-Verfahren $8,76 \pm 4,46$, für das *llm*-Verfahren $96,6 \pm 0,95$ und für das *json*-Verfahren $99,91 \pm 0,09$.

Durch Hinzunahme des Landes bei den Anfragen ändert sich nicht, dass das *regex*-Verfahren den größten Fehler und mit Abstand die wenigsten zulässigen Antworten produziert und die beiden anderen Verfahren relativ nah beisammen sind.

Auswirkungen der Hinzunahme der Länderinformation

Die Differenz zwischen den Ergebnissen des Experiments *Koordinaten* (3.1) ohne die Länderinformation (vgl. Abbildung 4.1) und mit der Länderinformation (vgl. Abbildung 4.3) ist in der Abbildung 4.5 dargestellt. Anhand der Abbildung lässt sich erkennen, dass die Länderinformation keine einheitliche und tendenziell eine verschlechternde Wirkung auf das *regex*-Verfahren hat. Bei den anderen beiden Verfahren werden die Ergebnisse durch die Hinzunahme des Landes für jedes Modell besser. Dabei fällt auf, dass die Verbesserung für das *json*-Verfahren bei jedem Modell größer ist, als beim *llm*-Verfahren. Das sorgt dafür, dass die Differenz der Fehler der *json*- und *llm*-Verfahren nach

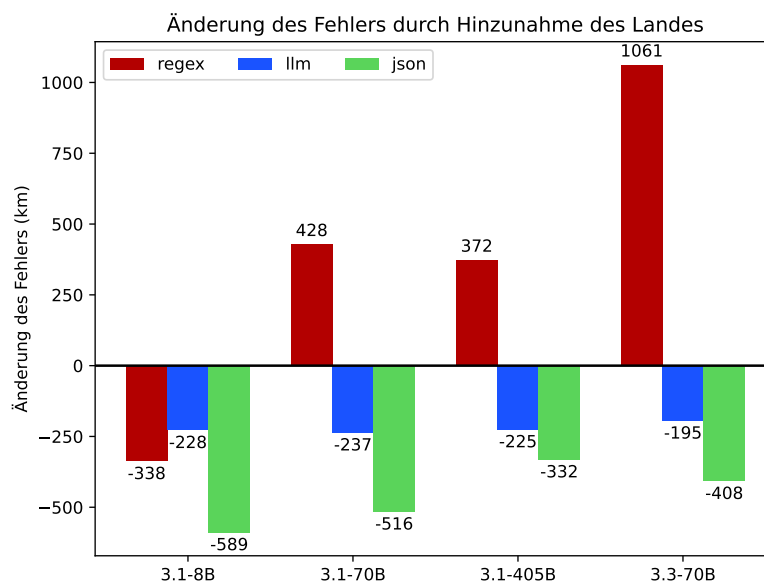


Abbildung 4.5: Die Differenz zwischen den in den Abbildungen 4.1 und 4.3 dargestellten Ergebnissen des Experiments *Koordinaten* (3.1). Dadurch wird der Einfluss der Länderinformation in den Anfragen dargestellt. Ein negativer Wert bedeutet in dem Fall, dass der Fehler durch die Hinzunahme der Länderinformation um diesen Wert kleiner geworden ist.

Hinzunahme der Länderinformation deutlich kleiner geworden ist (vgl. Abbildungen 4.1 und 4.3).

Standardabweichung

Allerdings fällt auf, dass die Standardabweichung für alle Modelle und alle Anfragearten sehr hoch ist und sich relativ zum durchschnittlichen Fehler durch die Hinzunahme der Länderinformation kaum ändert (vgl. Abbildung 4.6).

4.1.2 Qualitative Beispiele und Erörterung

Im Folgenden werden einzelne Anfragen qualitativ genauer betrachtet, um das Verhalten der LLMs bei den verschiedenen Anfragearten und häufige Fehler darzustellen.

Für die Unterkapitel *Ablauf und Probleme* der verschiedenen Verfahren wird das Llama-3.1-8B-Instruct Modell mit Anfragen für Abu Dhabi ohne die Länderinformation genutzt. Außerdem werden die in Kapitel 3 beschriebenen System Prompts und Vorlagen genutzt, um die Anfragen zu generieren.

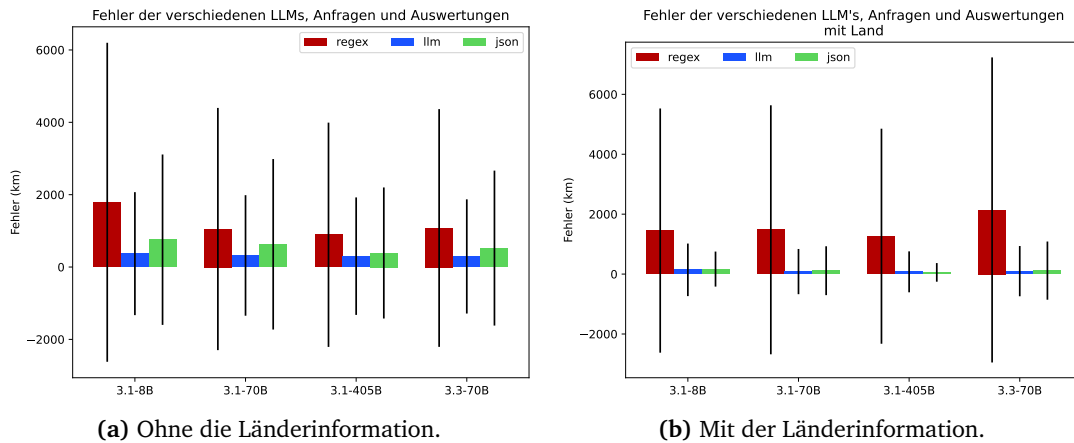


Abbildung 4.6: Der durchschnittliche Fehler von den verschiedenen LLMs, Anfragen und Auswertungsverfahren einmal mit und einmal ohne die Länderinformation für das Experiment *Koordinaten* (3.1). Dabei werden die Fehler für die LLMs Llama-3.1-Instruct in den Größen 8B, 70B, 405B und Llama-3.3-70B-Instruct (links nach rechts) für jeweils die Auswertungsverfahren *regex* (rot), *llm* (blau) und *json* (grün) mit den zugehörigen Vorlagen visualisiert. Zusätzlich wird außerdem die Standardabweichung für jeden Anfragetypen in schwarz visualisiert.

Ablauf und Probleme: *regex*-Verfahren

Bei dem *regex*-Verfahren wird eine Vorlage verwendet, die kein Antwortformat vorschreibt (vgl. Abbildung 3.3). Dadurch entsteht die längere Antwort 4.7. Trotz der Tatsache, dass die Antwort offensichtlich Koordinaten enthält, schafft es das *regex*-Verfahren nicht, Koordinaten zu extrahieren und damit einen Fehler zu berechnen. Das kommt sehr häufig vor².

Dies ist auch das Hauptproblem des *regex*-Verfahrens. Da aus sehr vielen Antworten fehlerhafterweise keine Koordinaten extrahiert werden, ist es nicht möglich, zuverlässige Werte zu erhalten.

Ablauf und Probleme: *llm*-Verfahren

Auch das *llm*-Verfahren wertet die Antwort 4.7 aus. Dafür wird, wie in Kapitel 3 beschrieben, eine Anfrage an das Llama-3.1-8B-Instruct-Turbo Modell mit dem System Prompt 3.4 und der vorigen Antwort 4.7 als Nutzereingabe gesendet. Dabei entsteht die Antwort 4.8. Diese Antwort kann sehr zuverlässig mit dem *json*-Verfahren ausgewertet werden.

²Da das *regex*- und *llm*-Verfahren dieselben Antworten auswerten, ist dies an dem Anteil der verarbeitbaren Antworten (s. Abbildung 4.2) erkennbar.

```
1      The coordinates of Abu Dhabi, the capital city of the United Arab
2      Emirates, are:
3
4      Latitude: 24.4653° N
5      Longitude: 54.3703° E
6
7      Please note that these coordinates are for the city center of Abu
8      Dhabi. If you need the coordinates for a specific location within
9      the city or the emirate, please let me know and I'll do my best
10     to provide you with the accurate information.
```

Abbildung 4.7: Die Antwort des Llama-3.1-8B-Instruct Modells auf die Anfrage für Abu Dhabi ohne die Länderinformation mit der Vorlage 3.3.

```
1      {"latitude": "24.4653", "longitude": "54.3703"}
```

Abbildung 4.8: Die Antwort des Llama-3.1-8B-Instruct-Turbo Modells auf die Anfrage des *llm*-Verfahrens für die vorige Antwort 4.7.

Dabei ist das Hauptproblem, dass beim Extrahieren der Koordinaten ein anderes LLM verwendet wird. Es kann nicht ausgeschlossen werden, dass dieses Modell fehlerhafte Koordinaten durch bessere ersetzt oder sogar bei Antworten, die keine Koordinaten enthalten, Koordinaten hinzufügt. Allerdings zeigen die durchschnittlichen Fehler für das *llm*-Verfahren (vgl. Abbildung 4.1), dass das Llama-3.1-8B-Instruct Modell den größten Fehler hat. Das lässt vermuten, dass das Llama-3.1-8B-Instruct-Turbo Modell, falls es nicht nur Koordinaten extrahieren sollte, das Ergebnis wahrscheinlich verschlechtert und nicht verbessert.

Ablauf und Probleme: *json*-Verfahren

Das *json*-Verfahren verwendet die Vorlage 3.2 und schreibt damit, genau wie die zweite Anfrage des *llm*-Verfahrens, ein festes Antwortformat vor. Dadurch entsteht die zu der Antwort 4.8 sehr ähnliche Antwort 4.9. Diese Antwort wird dann direkt ausgewertet (wie in Kapitel 3 beschrieben) und der Fehler berechnet.

Dabei gibt es keine häufig auftretenden Probleme, die mit dem Verfahren zusammenhängen.

```
1      {"latitude": 24.465646, "longitude": 54.369824}
```

Abbildung 4.9: Die Antwort des Llama-3.1-8B-Instruct Modells auf die Anfrage für Abu Dhabi ohne die Länderinformation mit der Vorlage 3.2.

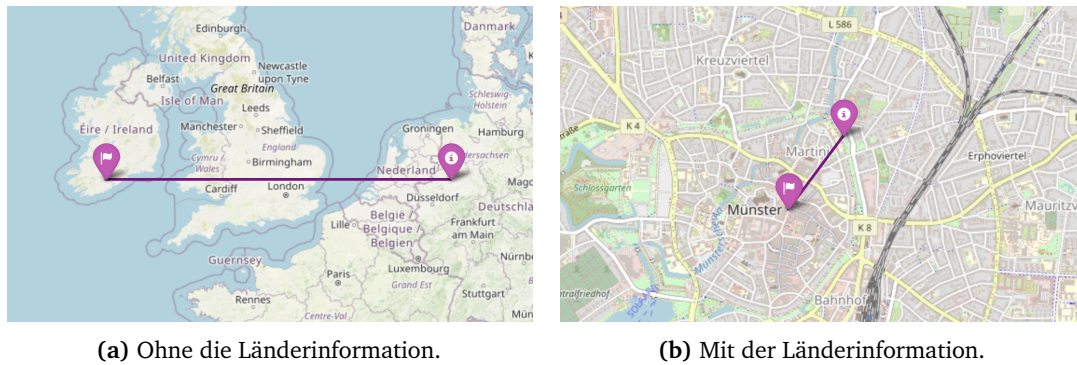


Abbildung 4.10: Die tatsächliche Position der Stadt Münster (NRW, Deutschland) und die Ergebnisse für das Experiment *Koordinaten* (3.1) von dem Llama-3.1-405B-Instruct Modell mit dem *json*-Verfahren. Dabei wird visualisiert, welchen Einfluss die Länderinformation für die zurückgegebenen Koordinaten von Münster hat.

Gleiche Städte- oder Regionennamen

Ein häufig auftretendes Problem ist, dass der Name einer Stadt oder Region nicht eindeutig ist, weil er z. B. in verschiedenen Ländern verwendet wird. Das führt zu besonders großen Fehlern, da das LLM nur anhand des Städtenamens nicht zwischen gleichnamigen Städten unterscheiden kann und eventuell die Koordinaten einer Stadt auf einem anderen Kontinent zurückgibt, die aber nicht gemeint ist.

In Abbildung 4.10 wird visualisiert, dass diese Verwechslung mit dem Llama-3.1-405B-Instruct Modell bei dem *json*-Verfahren für die Stadt Münster (NRW, Deutschland) passiert ist. Außerdem ist dargestellt, dass durch die Hinzunahme des Landes die Verwechslung verhindert wurde.

Allerdings werden dadurch nicht alle Verwechslungsprobleme gelöst. Es kann auch vorkommen, dass innerhalb eines Landes der gleiche Städtenamen mehrmals verwendet wird. Das ist für den Gesamtfehler vor allem bei großen Ländern problematisch. In der Abbildung 4.11 wird dargestellt, wie das Llama-3.1-8B-Instruct Modell mit dem *json*-Verfahren trotz der Länderinformation die Koordinaten für Pasadena, Kalifornien³ zurückgibt, obwohl Pasadena, Texas⁴ gemeint ist. Dadurch entsteht ein Fehler von über 2.200 km, der bei einem Durchschnittsfehler für diese Anfragen von ca. 167 km durchaus relevant ist (vgl. Abbildung 4.3).

³[https://de.wikipedia.org/wiki/Pasadena_\(Kalifornien\)](https://de.wikipedia.org/wiki/Pasadena_(Kalifornien))

⁴[https://de.wikipedia.org/wiki/Pasadena_\(Texas\)](https://de.wikipedia.org/wiki/Pasadena_(Texas))

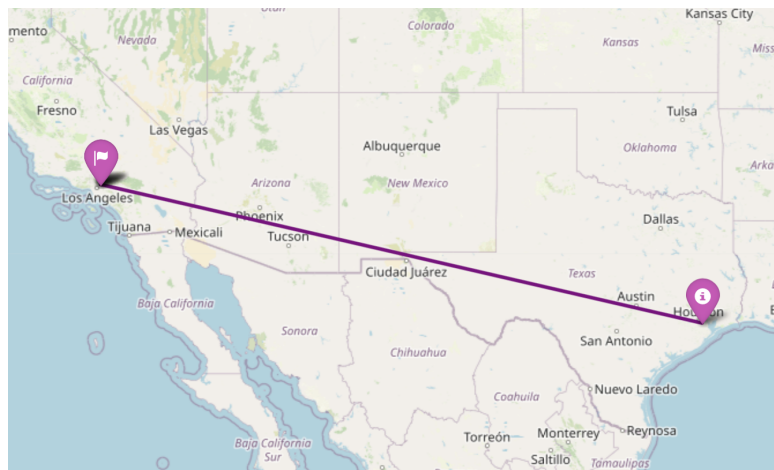


Abbildung 4.11: Das Llama-3.1-8B-Instruct Modell gibt mit dem *json*-Verfahren trotz der Länderinformation die Koordinaten für Pasadena, Kalifornien zurück, obwohl Pasadena, Texas gemeint ist.

4.2 Distanz

In diesem Abschnitt werden die Resultate des Experiments *Distanz* dargestellt. Diese basieren auf der in Kapitel 3.2 beschriebenen Vorgehensweise.

4.2.1 Quantitativer Vergleich der verschiedenen LLMs

Wie bei dem vorigen Experiment wird bei den Auswertungen und Visualisierungen in diesem Kapitel meistens der durchschnittliche Fehler genutzt. Damit ist der durchschnittliche Fehler aller Antworten gemeint, die die angegebenen Bedingungen erfüllen. Der einzelne Fehler wird, wie in Kapitel 3.2 beschrieben, berechnet.

Ohne die Länderinformation

Die durchschnittlichen Fehler ohne die Länderinformation zeigen: Je größer das LLM, desto besser das Ergebnis. Allerdings ist der Unterschied zwischen dem Llama-3.3-70B-Instruct Modell und dem Llama-3.1-405B-Instruct Modell sehr klein (vgl. Abbildung 4.12).

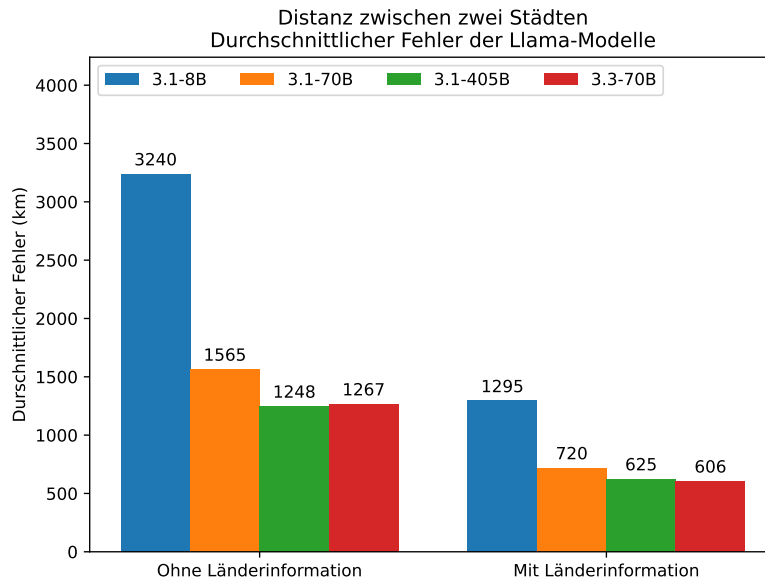


Abbildung 4.12: Der durchschnittliche Fehler von den verschiedenen LLMs mit und ohne die Länderinformation für das Experiment *Distanz* (3.2). Dabei werden die Fehler für die LLMs Llama-3.1-Instruct in den Größen 8B (blau), 70B (orange), 405B (grün) und Llama-3.3-70B-Instruct (rot) einmal mit (rechts) und einmal ohne (links) die Länderinformation visualisiert.

Mit der Länderinformation

Auch bei den durchschnittlichen Fehlern mit der Länderinformation zeigt sich, dass größere Modelle meistens bessere Ergebnisse produzieren (vgl. Abbildung 4.12). Allerdings liefert das Llama-3.3-70B-Instruct Modell die besten Ergebnisse, obwohl es nicht das größte Modell ist. Dabei muss dennoch beachtet werden, dass hier die älteren Llama 3.1 Modelle⁵ mit dem neueren Llama 3.3 Modell⁶ verglichen werden.

Auswirkungen der Hinzunahme der Länderinformation

In Abbildung 4.12 ist gut erkennbar, dass sich der Fehler durch die Hinzunahme der Länderinformation für alle Modelle ungefähr halbiert. Dazu passt, dass sich beim Experiment *Koordinaten* der Fehler für das *json*-Verfahren durch Hinzunahme der Länderinformation deutlich verbessert hat.

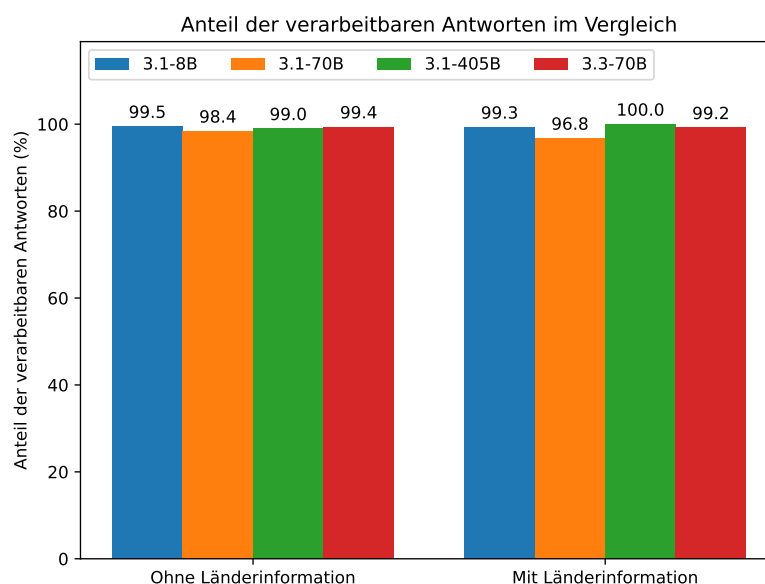


Abbildung 4.13: Der Anteil der korrekt verarbeitbaren Antworten der verschiedenen LLMs mit und ohne die Länderinformation für das Experiment *Distanz* (3.2). Dabei werden die Anteile der Antworten, die korrekt verarbeitet werden können, von den LLMs Llama-3.1-Instruct in den Größen 8B (blau), 70B (orange), 405B (grün) und Llama-3.3-70B-Instruct (rot) einmal mit (rechts) und einmal ohne (links) die Länderinformation visualisiert.

Anteil der verarbeitbaren Antworten

Der Anteil der verarbeitbaren Antworten ist für alle LLMs sehr groß und es gibt keine signifikanten Unterschiede (vgl. Abbildung 4.13). Die Mittelwerte betragen für die Antworten ohne die Länderinformation $99,06 \pm 0,42$ und für die Antworten mit der Länderinformation $98,82 \pm 1,23$. Diese Ergebnisse stimmen mit den nahezu perfekten Antwortraten des *json*-Verfahrens beim Experiment *Koordinaten* überein.

Standardabweichung

Allerdings fällt auch bei diesem Experiment auf, dass die Standardabweichung für alle Modelle sehr hoch ist. Die Hinzunahme der Länderinformation verringert die Standardabweichung deutlich. Trotzdem liegt die geringste Standardabweichung aller LLMs, die vom Llama-3.3-70B Modell erreicht wird, bei ca. 616 km.

⁵https://www.llama.com/docs/model-cards-and-prompt-formats/llama3_1/

⁶https://www.llama.com/docs/model-cards-and-prompt-formats/llama3_3/

```

1      [
2          {
3              'role': 'system',
4              'content': 'Answer in json format only: {"distance": < e.g.
                          '"1945.399479"' >}. The value must be in kilometers and
                          no explanation, just the answer!'
5          },
6          {
7              'role': 'user',
8              'content': 'What is the distance between bharuch, India and
                          berlin, Germany?'
9          }
10     ]

```

Abbildung 4.14: Die Anfrage für Bharuch, Indien und Berlin, Deutschland mit der Vorlage 3.5 unter Verwendung der Länderinformation.

```

1      {
2          "distance": 6171.479892
3      }

```

Abbildung 4.15: Die Antwort des Llama-3.1-8B-Instruct Modells für die Anfrage 4.14.

4.2.2 Qualitative Beispiele und Erörterung

Im Folgenden wird eine einzelne Anfrage etwas genauer vorgestellt und auf Probleme eingegangen, um die Ergebnisse und die Vorgehensweise nachvollziehbarer zu machen.

Beispiel-Anfrage

Es wird die Anfrage für Bharuch, Indien und Berlin, Deutschland betrachtet. Mithilfe der Vorlage 3.5 und der Länderinformation wird die Anfrage 4.14 erstellt. Die Antwort des Llama-3.1-8B-Instruct Modells auf diese Anfrage ist in der Abbildung 4.15 dargestellt und entspricht einer Standardantwort für das *json*-Verfahren. Mit dieser Antwort kann die Distanz direkt extrahiert werden. In diesem Beispiel beträgt der Fehler ca. 100 km.

Probleme

Die meisten Probleme, die bei dem Experiment *Koordinaten* auftreten, werden hier direkt durch den Gebrauch des *json*-Verfahrens vermieden. Dadurch gibt es kaum unzulässige Antworten und es besteht nicht die Möglichkeit, dass im Laufe der Auswertung die Informationen verändert werden.

Allerdings gibt es auch hier das Problem, dass die Städtenamen nicht eindeutig sind. Das wird teilweise durch die Hinzunahme der Länderinformation behoben. Trotzdem gibt es weiterhin das Problem gleichnamiger Städte in einem Land.

4.3 Schlüsselergebnisse beider Experimente

Bei dem Experiment *Distanz* zeigen sich ähnliche Ergebnisse, wie bei dem Experiment *Koordinaten*:

1. **json-Verfahren.** Das *json*-Verfahren zeigt in beiden Experimenten nahezu perfekte Antwortraten – unabhängig von der Länderinformation (vgl. Abbildungen 4.2, 4.4 und 4.13).
2. **Modellgröße.** In beiden Experimenten gilt, bis auf wenige Ausnahmen: Je größer das LLM, desto besser die Ergebnisse (vgl. Abbildungen 4.1, 4.3 und 4.12).
3. **Länderinformation.** In beiden Experimenten verbessert die Hinzunahme der Länderinformation die Ergebnisse⁷ (vgl. Abbildungen 4.5 und 4.12).

Außerdem zeigen sich bei dem Experiment *Koordinaten* folgende Ergebnisse:

4. **regex-Verfahren.** Das *regex*-Verfahren erzeugt unabhängig von der Länderinformation kaum auswertbare Antworten (vgl. Abbildungen 4.2 und 4.4).
5. **llm-Verfahren.** Das *llm*-Verfahren liefert für alle LLMs, bis auf eine einzige Ausnahme, unabhängig von der Länderinformation die geringsten durchschnittlichen Fehler (vgl. Abbildungen 4.1 und 4.3) und erzeugt kaum nicht auswertbare Antworten (vgl. Abbildungen 4.2 und 4.4).

⁷Ausnahme ist das *regex*-Verfahren beim Experiment *Koordinaten*. Allerdings sind bei diesem Verfahren die Anteile der zulässigen Antworten so gering, dass keine zuverlässigen Werte ermittelt werden können.

5 | Diskussion und Fazit

Im Folgenden werden die in Kapitel 3 beschriebene Methodik und die in Kapitel 4 präsentierten Ergebnisse kritisch reflektiert. Dabei werden sowohl die Einschränkungen, Stärken und Schwächen als auch mögliche Ansätze für weiterführende Arbeiten herausgearbeitet.

Anschließend werden die Ergebnisse zusammengefasst und hinsichtlich möglicher praktischer Einsatzmöglichkeiten diskutiert. Zudem wird erörtert, ob damit die folgende Fragestellung beantwortet werden kann:

In welchem Umfang sind ausgewählte Llama 3 Modelle in der Lage, akkurate Koordinaten für Städte anzugeben und Distanzen zwischen Städten zu bestimmen?

Abschließend werden die Bedeutung der Ergebnisse herausgearbeitet und ein Fazit gezogen.

5.1 Diskussion und Perspektiven für zukünftige Forschung

Zuerst werden allgemeine Einschränkungen, Stärken, Schwächen und weiteres Forschungspotential dargestellt. Anschließend wird dies auch für die beiden Experimente *Koordinaten* und *Distanz* gemacht.

Ein zentraler Beitrag dieser Arbeit ist der Vergleich von vier verschiedenen Llama 3 Modellen, da dadurch sowohl die aktuellen Unterschiede dargestellt werden, als auch eine Grundlage für den Vergleich mit zukünftigen Modellen geschaffen wird. Weitere Arbeiten könnten die selben Experimente mit anderen LLMs oder neueren Llama Modellen durchführen, um Unterschiede bzw. Fortschritte festzustellen.

Allerdings liegt eine Einschränkung dieser Arbeit in der Auswahl der Daten, da nur Städte mit mindestens 100.000 Einwohnern betrachtet werden. Daher kann nur spekuliert werden, wie sich die Modelle und Verfahren für kleinere Städte verhalten. In zukünftigen

Model	Error (km)	P-Rate (%)
LLaMA (7B)	521	10
LLaMA (13B)	386	31
Llama 3.1 (8B)	1790	12
Llama 3.1 (70B)	1051	8
Llama 3.1 (405B)	892	12
Llama 3.3 (70B)	1079	4

Tabelle 5.1: Vergleich der Ergebnisse von Bhandari et al. (2023) (oben) mit den Ergebnissen des *regex*-Verfahren (unten) ohne die Länderinformation (s. Abbildungen 4.1 und 4.2). Dabei ist die *P-Rate* der prozentuale Anteil der Antworten, die erfolgreich ausgewertet werden können.

Arbeiten könnte untersucht werden, wie sich LLMs bei kleineren Städten oder anderen geographischen Punkten verhalten.

Außerdem konnte durch die experimentelle Evaluation bestätigt werden, dass die Auswahl der Daten, die in den Anfragen an die LLMs enthalten sind, essentiell für die Genauigkeit der Antworten ist. In dieser Arbeit wurde beispielhaft gezeigt, dass die Länderinformation einen sehr großen Einfluss auf die Antworten hat. Die Auswahl dieser Daten bietet weiteres Forschungspotential, beispielsweise den Einfluss von der Ergänzung der Bevölkerungszahl oder des Bundesstaates/-landes zum Städtenamen auf die Genauigkeit.

Koordinaten

Eine Stärke dieser Arbeit liegt in der Analyse von einem bereits bekannten und zwei neuen Auswertungsverfahren in Kombination mit den vier verschiedenen Llama Modellen. Durch diese breite Betrachtung wird eine differenzierte Bewertung der Fähigkeiten der Modelle und der Einflüsse der Auswertungsverfahren ermöglicht. In Zukunft könnten noch weitere Auswertungsverfahren betrachtet werden, z. B. ein aktualisiertes *regex*-Verfahren.

Das *regex*-Verfahren von Bhandari et al. (2023) wurde unverändert übernommen, um einen direkten Vergleich der LLMs zu ermöglichen. Die Ergebnisse der Arbeit werden in der Abbildung 5.1 mit den Ergebnissen des *regex*-Verfahrens ohne die Länderinformation dieses Experimentes verglichen. Dabei fällt auf, dass bei beiden Arbeiten größere Modelle kleinere Fehler aufweisen. Allerdings sind die bereits geringen *P-Rate*-Werte vor allem bei dem neusten Modell (Llama 3.3) deutlich gesunken.

Eine Einschränkung dieser Arbeit liegt in der fehlenden Untersuchung dieser Veränderungen. Die Frage, warum bei neueren Modellen die *P-Rate* und der Fehler schlechter sind, konnte nicht eindeutig beantwortet werden, da es aufgrund der fehlenden Verfügbarkeit nicht möglich war, direkt mit dem LLaMa Modell zu arbeiten. Eine mögliche Erklärung wäre, dass die neueren Modelle in einem anderen Format antworten, das von den regulären Ausdrücken nicht erkannt wird.

Allerdings zeigen die erzielten Ergebnisse, dass die *json*- und *llm*-Verfahren vergleichbar gute und teilweise sogar bessere Ergebnisse produzieren, als Bhandari et al. (2023) es mit dem *regex*-Verfahren erreicht haben. Ein wesentlicher Vorteil dieser Methoden sind die verglichen mit dem *regex*-Verfahren nahezu perfekten *P-Rate* Werte.

Ein möglicher Nachteil des *llm*-Verfahrens ist, dass nicht ausgeschlossen werden kann, dass bei der Extraktion der Koordinaten durch ein LLM (vgl. Kapitel 3.1.4) Informationen verändert und nicht nur extrahiert werden. Dies könnte in anschließenden Arbeiten untersucht werden, da die Fähigkeit, einzelne Informationen aus unstrukturierten Texten zu extrahieren, vielseitig eingesetzt werden könnte.

Distanz

Ein möglicher Nachteil dieses Experimentes ist, dass nur das *json*-Verfahren betrachtet wurde. Es wurde aufgrund der Ergebnisse des vorigen Experimentes ausgewählt. Allerdings besteht die Möglichkeit, dass z. B. das *llm*- oder ein neues Verfahren deutlich bessere Ergebnisse liefern würde. In Zukunft könnten für diesen Anwendungsfall noch weitere Auswertungsverfahren getestet werden.

5.2 Zusammenfassung und Fazit

Das Experiment *Koordinaten* (s. Kapitel 3.1 und 4.1) sollte beantworten, *in welchem Umfang ausgewählte Llama 3 Modelle in der Lage sind, akkurate Koordinaten für Städte anzugeben.*

Die Ergebnisse des Experimentes legen nahe, dass die Antwort stark vom Kontext abhängt. Während der durchschnittliche Fehler mit Länderinformation bei manchen Modellen mit ausgewählten Verfahren unter 100 km liegt und damit im globalen Kontext als akzeptabel betrachtet werden könnte, sind die Fehler ohne die Länderinformation signifikant größer und überschreiten immer 300 km. Zudem zeigt die hohe Standardabweichung, dass

einzelne Antworten keine akkuraten Koordinaten für Städte zurückgeben - auch nach Optimierungen bleibt das Problem bestehen.

Besonders das *regex*-Verfahren zeigt sich als praktisch unbrauchbar im Vergleich zu den anderen Verfahren, da nur ein sehr geringer Anteil der Antworten korrekt ausgewertet werden kann. Obwohl die verschiedenen Verfahren bei bestimmten Anfragen, vor allem mit der Länderinformation, vielversprechende Ergebnisse liefern, zeigen die hohen Standardabweichungen, dass diese Verfahren für den praktischen Einsatz noch ungeeignet sind, da bei allen potentiellen Einsatzgebieten nicht kleine durchschnittliche Fehler, sondern akkurate einzelne Werte benötigt werden.

Das Experiment *Distanz* (s. Kapitel 3.2 und 4.2) sollte beantworten, *in welchem Umfang ausgewählte Llama 3 Modelle in der Lage sind, Distanzen zwischen Städten zu bestimmen.*

Die Ergebnisse des Experimentes zeigen, dass das hier verwendete *json*-Verfahren selbst mit der Länderinformation bei keinem der betrachteten Modelle einen durchschnittlichen Fehler von unter 600 km erreicht hat. Außerdem beträgt die minimale Standardabweichung über 600 km. Demnach ist es mit dieser Methode und diesen Modellen nicht möglich, kleine durchschnittliche Fehler zu erreichen oder akkurate Distanzen zwischen einzelnen Städten zu bestimmen.

Die Ergebnisse dieser Arbeit legen nahe, dass größere und neuere Modelle ein besseres geographisches Wissen haben als kleinere und ältere Modelle. Abschließend bedeutet das für die drei in der Einleitung erwähnten Punkte Sicherheit, Fortschritt und Anwendungsmöglichkeiten, dass ein Fortschritt stattfindet, aber die Leistung noch nicht gut genug für praktische Anwendungsmöglichkeiten ist und man sich aufgrund der großen Standardabweichung nicht auf die Antworten verlassen darf.

Literaturverzeichnis

- AI @ Meta, L. T. (2024). The Llama 3 Herd of Models | Research - AI at Meta.
<https://ai.meta.com/research/publications/the-llama-3-herd-of-models/>.
- Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete Problems in AI Safety.
- Azerbayev, Z., Schoelkopf, H., Paster, K., Santos, M. D., McAleer, S., Jiang, A. Q., Deng, J., Biderman, S., & Welleck, S. (2024). Llemma: An Open Language Model For Mathematics.
- Bhandari, P., Anastasopoulos, A., & Pfoser, D. (2023). Are Large Language Models Geospatially Knowledgeable? In *Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems, SIGSPATIAL '23*, (pp. 1–4). New York, NY, USA: Association for Computing Machinery.
- Capgemini (2023). Generative KI - Verwendungszweck 2023.
<https://de.statista.com/statistik/daten/studie/1403840/umfrage/verwendungszweck-generativer-ki-tools/>.
- Horizont (2023). Umfrage zur Nutzung von KI-Diensten 2023.
<https://de.statista.com/statistik/daten/studie/1373267/umfrage/umfrage-zur-aktiven-nutzung-von-ki-diensten-in-deutschland/>.
- Karney, C. F. F. (2013). Algorithms for geodesics. *Journal of Geodesy*, 87(1), 43–55.
- Li, Y., Li, Z., Zhang, K., Dan, R., Jiang, S., & Zhang, Y. (2023). ChatDoctor: A Medical Chat Model Fine-Tuned on a Large Language Model Meta-AI (LLaMA) Using Medical Domain Knowledge. *Cureus*, 15(6), e40895.
- OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., Bello, I., Berdine, J., Bernadett-Shapiro, G., Berner, C., Bogdonoff, L., Boiko, O., Boyd, M., Brakman, A.-L., Brockman, G., Brooks, T., Brundage, M., Button, K., Cai, T., Campbell, R., Cann, A., Carey, B.,

- Carlson, C., Carmichael, R., Chan, B., Chang, C., Chantzis, F., Chen, D., Chen, S., Chen, R., Chen, J., Chen, M., Chess, B., Cho, C., Chu, C., Chung, H. W., Cummings, D., Currier, J., Dai, Y., Decareaux, C., Degry, T., Deutsch, N., Deville, D., Dhar, A., Dohan, D., Dowling, S., Dunning, S., Ecoffet, A., Eleti, A., Eloundou, T., Farhi, D., Fedus, L., Felix, N., Fishman, S. P., Forte, J., Fulford, I., Gao, L., Georges, E., Gibson, C., Goel, V., Gogineni, T., Goh, G., Gontijo-Lopes, R., Gordon, J., Grafstein, M., Gray, S., Greene, R., Gross, J., Gu, S. S., Guo, Y., Hallacy, C., Han, J., Harris, J., He, Y., Heaton, M., Heidecke, J., Hesse, C., Hickey, A., Hickey, W., Hoeschele, P., Houghton, B., Hsu, K., Hu, S., Hu, X., Huizinga, J., Jain, S., Jain, S., Jang, J., Jiang, A., Jiang, R., Jin, H., Jin, D., Jomoto, S., Jonn, B., Jun, H., Kaftan, T., Kaiser, Ł., Kamali, A., Kanitscheider, I., Keskar, N. S., Khan, T., Kilpatrick, L., Kim, J. W., Kim, C., Kim, Y., Kirchner, J. H., Kiros, J., Knight, M., Kokotajlo, D., Kondraciuk, Ł., Kondrich, A., Konstantinidis, A., Kopic, K., Krueger, G., Kuo, V., Lampe, M., Lan, I., Lee, T., Leike, J., Leung, J., Levy, D., Li, C. M., Lim, R., Lin, M., Lin, S., Litwin, M., Lopez, T., Lowe, R., Lue, P., Makanju, A., Malfacini, K., Manning, S., Markov, T., Markovski, Y., Martin, B., Mayer, K., Mayne, A., McGrew, B., McKinney, S. M., McLeavey, C., McMillan, P., McNeil, J., Medina, D., Mehta, A., Menick, J., Metz, L., Mishchenko, A., Mishkin, P., Monaco, V., Morikawa, E., Mossing, D., Mu, T., Murati, M., Murk, O., Mély, D., Nair, A., Nakano, R., Nayak, R., Neelakantan, A., Ngo, R., Noh, H., Ouyang, L., O’Keefe, C., Pachocki, J., Paino, A., Palermo, J., Pantuliano, A., Parascandolo, G., Parish, J., Parparita, E., Passos, A., Pavlov, M., Peng, A., Perelman, A., Peres, F. d. A. B., Petrov, M., Pinto, H. P. d. O., Michael, Pokorny, Pokrass, M., Pong, V. H., Powell, T., Power, A., Power, B., Proehl, E., Puri, R., Radford, A., Rae, J., Ramesh, A., Raymond, C., Real, F., Rimbach, K., Ross, C., Rotsted, B., Roussez, H., Ryder, N., Saltarelli, M., Sanders, T., Santurkar, S., Sastry, G., Schmidt, H., Schnurr, D., Schulman, J., Selsam, D., Sheppard, K., Sherbakov, T., Shieh, J., Shoker, S., Shyam, P., Sidor, S., Sigler, E., Simens, M., Sitkin, J., Slama, K., Sohl, I., Sokolowsky, B., Song, Y., Staudacher, N., Such, F. P., Summers, N., Sutskever, I., Tang, J., Tezak, N., Thompson, M. B., Tillet, P., Tootoonchian, A., Tseng, E., Tuggle, P., Turley, N., Tworek, J., Uribe, J. F. C., Vallone, A., Vijayvergiya, A., Voss, C., Wainwright, C., Wang, J. J., Wang, A., Wang, B., Ward, J., Wei, J., Weinmann, C. J., Welihinda, A., Welinder, P., Weng, J., Weng, L., Wiethoff, M., Willner, D., Winter, C., Wolrich, S., Wong, H., Workman, L., Wu, S., Wu, J., Wu, M., Xiao, K., Xu, T., Yoo, S., Yu, K., Yuan, Q., Zaremba, W., Zellers, R., Zhang, C., Zhang, M., Zhao, S., Zheng, T., Zhuang, J., Zhuk, W., & Zoph, B. (2023). GPT-4 Technical Report.
- Roberts, J., Lüddecke, T., Das, S., Han, K., & Albanie, S. (2023). GPT4GEO: How a Language Model Sees the World’s Geography.
- Rozière, B., Gehring, J., Gloeckle, F., Sootla, S., Gat, I., Tan, X. E., Adi, Y., Liu, J., Sauvestre, R., Remez, T., Rapin, J., Kozhevnikov, A., Evtimov, I., Bitton, J., Bhatt, M.,

- Ferrer, C. C., Grattafiori, A., Xiong, W., Défossez, A., Copet, J., Azhar, F., Touvron, H., Martin, L., Usunier, N., Scialom, T., & Synnaeve, G. (2024). Code Llama: Open Foundation Models for Code.
- Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023a). LLaMA: Open and Efficient Foundation Language Models.
- Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P. S., Lachaux, M.-A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E. M., Subramanian, R., Tan, X. E., Tang, B., Taylor, R., Williams, A., Kuan, J. X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., & Scialom, T. (2023b). Llama 2: Open Foundation and Fine-Tuned Chat Models.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All you Need. In *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc.
- Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., Liu, P., Nie, J.-Y., & Wen, J.-R. (2024). A Survey of Large Language Models.

Hilfsmittel

Für diese Arbeit wurden verschiedene Hilfsmittel eingesetzt, um sowohl sprachliche als auch strukturelle Aspekte zu optimieren. Hier wird ein kurzer Überblick über die wichtigsten verwendeten Hilfsmittel gegeben:

1. **DeepL**¹ wurde hauptsächlich bei der Literaturrecherche zur Übersetzung vom Englischen ins Deutsche verwendet und zur Suche nach deutschen Formulierungen für englische Ausdrücke.
2. **DeepL Write**² wurde als Formulierungshilfe genutzt. Dabei war besonders die Funktion *Wort austauschen* hilfreich, um passende Synonyme bei Wortwiederholungen zu finden.
3. **ChatGPT**³ wurde als nützliches Hilfsmittel zum Brainstormen verwendet. Dabei wurde zum Beispiel über viele Versionen des Titels und der Fragestellung dieser Arbeit iteriert. Außerdem wurden im Rahmen eines Dialoges mögliche Kapitelaufteilungen erstellt, die tatsächliche Ausformulierung ist aber immer selbständig erfolgt.

¹<https://www.deepl.com/de/translator>

²<https://www.deepl.com/de/write>

³<https://chatgpt.com/>

Eigenständigkeitserklärung

Hiermit versichere ich, dass die vorliegende Arbeit über

*Evaluierung der geographischen Fähigkeiten ausgewählter Llama 3 Modelle:
Eine Untersuchung zur Genauigkeit der Koordinaten- und Distanzermittlung von Städten*

selbstständig von mir und ohne fremde Hilfe verfasst worden ist, dass keine anderen Quellen und Hilfsmittel, als die angegebenen benutzt worden sind und dass die Stellen der Arbeit, die anderen Werken – auch elektronischen Medien und KI-Tools – dem Wortlaut oder Sinn nach entnommen wurden, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht worden sind. Mir ist bekannt, dass es sich bei einem Plagiat um eine Täuschung handelt, die gemäß der Prüfungsordnung sanktioniert werden kann.

Ich erkläre hiermit, dass ich Kenntnis von einer zum Zweck der Plagiatskontrolle vorzunehmenden Speicherung der Arbeit in einer Datenbank sowie von ihrem Abgleich mit anderen Texten zwecks Auffindung von Übereinstimmungen habe.

Ich versichere, dass ich die vorliegende Arbeit oder Teile daraus nicht anderweitig als Prüfungsarbeit eingereicht habe.

(Ort, Datum)

(Unterschrift)