# Auswahl der Bewertungsmetriken

Für die Evaluation der Performance der verschiedenen LLMs (Llama, Gemma, Mistral) im Vergleich zu den gegebenen Referenzantworten wurden drei Metriken ausgewählt, die auf die spezifischen Anforderungen des Usecases abgestimmt sind. Dieser Usecase umfasst die Evaluierung von Modellantworten auf Fragen zu aktuellen NBA-Ereignissen und -Leistungen, wobei neben der sprachlichen Präzision auch die faktische Korrektheit und die semantische Übereinstimmung im Fokus stehen. Die Auswahl der Metriken basiert auf den folgenden Überlegungen:

## **ROUGE-N (n-Gramm-Überlappung)**

ROUGE-N misst die Überlappung von n-Grammen (d.h. Sequenzen von n benachbarten Wörtern) zwischen den Modellantworten und den Referenzantworten. Diese Metrik ist besonders geeignet, da sie die Präzision der Wortwahl und der syntaktischen Struktur beurteilt. Insbesondere bei der Beantwortung von Fragen zu sportlichen Ereignissen wie Spielergebnissen oder Teamstatistiken ist es entscheidend, dass die Modelle die relevanten Begriffe korrekt und präzise wiedergeben. ROUGE-N ermöglicht eine quantitative Messung, wie genau das Modell wichtige n-Gramme übernommen hat, was für die Beurteilung der Genauigkeit der Modellantworten wichtig ist.

Innerhalb von ROUGE-N werden mehrere Metriken verwendet, die jeweils unterschiedliche Aspekte der n-Gramm-Überlappung messen:

* + **ROUGE-1** misst die Übereinstimmung der einzelnen Wörter (Unigramme).
  + **ROUGE-2** prüft die Übereinstimmung der Paarwörter (Bigramme).
  + **ROUGE-L** berücksichtigt die längste gemeinsame Teilsequenz (Longest Common Subsequence), um die allgemeine Struktur der Antwort zu bewerten.
  + **ROUGE-Lsum** misst die Übereinstimmung für die Zusammenfassung der Antwort, wobei die längste gemeinsame Teilsequenz als Maßstab verwendet wird.

Die Wahl dieser vier Metriken basiert darauf, dass sie unterschiedliche Dimensionen der Wortwahl und syntaktischen Struktur abdecken. Während ROUGE-1 grundlegende WortÜbereinstimmungen prüft, ergänzt ROUGE-2 die Analyse mit der Untersuchung von Wortpaaren. ROUGE-L und ROUGE-Lsum bieten eine tiefergehende Analyse der Struktur und Kohärenz der Antwort. Um eine umfassende Beurteilung zu gewährleisten, wird der F1-Score aus den Precision- und Recall-Werten dieser Metriken verwendet, da er sowohl die Genauigkeit als auch die Vollständigkeit der Übereinstimmung zwischen Modellantwort und Referenz widerspiegelt. Der F1-Score bietet somit eine ausgewogene Metrik, die Verzerrungen durch extreme Precision- oder Recall-Werte vermeidet und eine präzisere Einschätzung der Modellleistung ermöglicht.

## **BERTScore (Semantische Ähnlichkeit)**

BERTScore nutzt die kontextuellen Wortdarstellungen aus BERT, um die semantische Ähnlichkeit zwischen Modellantworten und Referenzantworten zu berechnen. Diese Metrik ist besonders wichtig, wenn Modelle unterschiedliche Formulierungen oder Synonyme verwenden, die jedoch denselben Informationsgehalt haben. BERTScore stellt sicher, dass die semantische Bedeutung der Antwort mit der Referenz übereinstimmt, auch wenn sich die Wortwahl unterscheidet.

Für die Bewertung von Antworten auf Fragen zu aktuellen Ereignissen in der NBA ist es entscheidend, dass das Modell die wesentlichen Informationen korrekt wiedergibt, auch wenn es alternative Formulierungen verwendet. BERTScore eignet sich daher hervorragend zur Messung der Kohärenz der Modellantworten und ihrer semantischen Ähnlichkeit mit den Referenzantworten.

Im Rahmen dieser Evaluation wird der F1-Score des BERTScore verwendet, um die semantische Übereinstimmung zwischen den Modellantworten und den Referenzantworten zu messen. Der F1-Score kombiniert Precision und Recall und bietet eine ausgewogene Metrik, die sowohl die Genauigkeit der semantischen Übereinstimmung als auch die Vollständigkeit der relevanten Informationen berücksichtigt. Durch den F1-Score wird sichergestellt, dass Modelle, die unterschiedliche Formulierungen verwenden, aber denselben Informationsgehalt korrekt wiedergeben, gut bewertet werden. Diese Metrik ermöglicht einen klaren Vergleich der Performance der Modelle (Llama, Gemma, Mistral) in Bezug auf ihre semantische Kohärenz und Genauigkeit.

## **FactCC (Fachliche Konsistenz)**

FactCC bewertet die faktische Korrektheit der Modellantworten und stellt sicher, dass die bereitgestellten Informationen mit den tatsächlichen Daten übereinstimmen. Im Kontext dieses Usecases, der auf aktuellen und verifizierbaren Daten basiert, ist es von zentraler Bedeutung, dass die Modellantworten keine fehlerhaften Informationen liefern. FactCC stellt sicher, dass alle Fakten in den Antworten korrekt sind und gibt eine verlässliche Einschätzung darüber, ob die Antwort den realen Ereignissen entspricht.

Diese Evaluierung kann entweder manuell durchgeführt werden, indem Modellantworten mit verifizierbaren Daten (z. B. NBA-Statistiken oder -Spielergebnissen) verglichen werden, oder automatisiert unter Verwendung von Datenbanken oder APIs, die aktuelle und verlässliche Informationen bereitstellen. In einer automatisierten Lösung wird beispielsweise eine API verwendet, die NBA-Daten liefert, um zu prüfen, ob die in der Modellantwort enthaltenen Informationen den realen Ergebnissen entsprechen. Eine korrekte Antwort wird mit der tatsächlichen Information abgeglichen und bewertet.

Zusammengefasst dient FactCC dazu, sicherzustellen, dass das Modell nicht nur sprachlich korrekt ist, sondern auch die richtigen und überprüfbaren Informationen liefert. Die Antwort wird als korrekt bewertet, wenn alle angegebenen Fakten mit den tatsächlichen Daten übereinstimmen.

In diesem Fall wurde die Evaluierung manuell durchgeführt, durch einen faktischen Vergleich mit den Referenzantworten. Bewertet wurde nicht binär sondern mithilfe einer geschätzten prozentuellen Übereinstimmung.

## **Kombinierte Bewertungsmethode: Gewichtung durch FactCC**

Diese Methode wurde gewählt, um die faktische Korrektheit (FactCC) als Grundlage der Bewertung zu priorisieren, während sprachliche Präzision (ROUGE) und semantische Kohärenz (BERTScore) ergänzend einbezogen werden. Der Ansatz stellt sicher, dass nur solche Antworten eine hohe Bewertung erhalten, die sowohl inhaltlich korrekt als auch sprachlich und semantisch überzeugend sind.

**Hintergrund der Entscheidung:** Faktentreue ist im gegebenen Usecase besonders entscheidend, da die Modellantworten auf aktuellen und überprüfbaren Daten basieren müssen. Eine sprachlich oder semantisch gute Antwort verliert ihren Wert, wenn die zugrunde liegenden Fakten nicht korrekt sind. Daher wird FactCC als Multiplikator genutzt, um sicherzustellen, dass sprachliche und semantische Bewertungen nur bei faktischer Korrektheit ins Gewicht fallen.

**Funktionsweise:** FactCC wird als Gewichtungsfaktor in die Berechnung der finalen Bewertung integriert. Dies geschieht, indem die Ergebnisse der ROUGE- und BERTScore-Metriken mit dem FactCC-Wert multipliziert werden. Der kombinierte Score ergibt sich aus dem Durchschnitt der gewichteten Werte. Diese Struktur gewährleistet eine klare und nachvollziehbare Bewertung, die sowohl die faktische als auch die sprachlich-semantische Qualität berücksichtigt.

**Vorteile der Methode:**

* + **Faktische Richtigkeit als Priorität:** Antworten mit niedriger oder fehlender Faktentreue erhalten keine hohe Bewertung, unabhängig von ihrer sprachlichen Qualität.
  + **Flexibilität und Ausgewogenheit:** Die Methode kombiniert mehrere Aspekte der Antwortqualität, ohne eine Dimension zu überbetonen.
  + **Nachvollziehbarkeit:** Die mathematische Struktur ermöglicht eine klare Interpretation und Vergleichbarkeit der Ergebnisse.

Zusammengefasst deckt diese kombinierte Bewertungsmethode die wichtigsten Aspekte der Antwortqualität ab, indem sie die faktische Richtigkeit priorisiert und gleichzeitig sprachliche und semantische Qualität einbezieht.

## **Vermeidung von Kontextverhaftung während der Tests:**

Während der ersten Testdurchläufe mit den 11 Fragen an die Modelle (Llama, Gemma, Mistral) wurde festgestellt, dass die Modelle dazu neigen, sich innerhalb eines kontinuierlichen Konversationskontexts zu stark auf zuvor genannte Spieler oder Themen zu fokussieren. Insbesondere beobachteten wir, dass die Modelle nach mehreren aufeinanderfolgenden Fragen über einen bestimmten Spieler weiterhin auf diesen Bezug nahmen, selbst wenn nachfolgende Fragen explizit andere Spieler thematisierten. Dies führte zu inhaltlichen Abweichungen und einer Beeinträchtigung der faktischen Konsistenz.

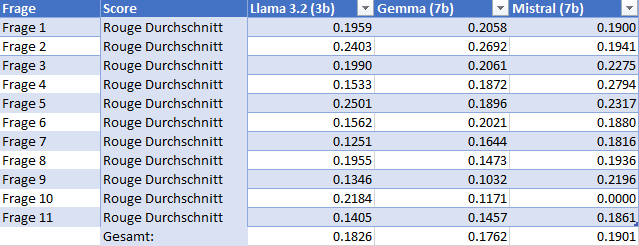
Um dieses Phänomen zu adressieren und die Evaluierung der Modellantworten objektiver zu gestalten, wurde ein zweiter Testdurchlauf durchgeführt. Dabei wurde vor jeder Frage zu einem neuen Spieler der Konversationskontext explizit zurückgesetzt, um sicherzustellen, dass die Modelle sich ausschließlich auf die jeweilige Frage und deren Inhalte fokussieren. Diese Maßnahme verhinderte, dass frühere Konversationen die Antworten beeinflussten, und ermöglichte eine präzisere Bewertung der Modellantworten unter kontrollierten Bedingungen.

# **Ergebnisse:**

Die vollständigen Ergebnisse können im Anhang in der Datei Evaluierung.xlsx eingesehen werden.

## **Ergebisse Testlauf 1:**

## **Rouge:**



In der Kategorie ROUGE, in der die sprachliche Präzision und strukturelle Kohärenz der Modellantworten gemessen wurde, erzielte Mistral den höchsten Durchschnittswert mit einem F1-Score von 0.1901, während Gemma mit einem Wert von 0.176 den niedrigsten Score aufwies. Der Durchschnittswert wurde berechnet, indem die F1-Scores von ROUGE-1, ROUGE-2, ROUGE-L und ROUGE-Lsum pro Modell für jede Frage gemittelt wurden. Anschließend ergab der Mittelwert über alle Fragen den Gesamtwert der jeweiligen Modelle in dieser Kategorie.

Obwohl Mistral in dieser Kategorie als Sieger hervorgeht, weisen alle drei Modelle insgesamt relativ niedrige Werte auf. Dies deutet darauf hin, dass die sprachliche Präzision und Kohärenz der Antworten im ersten Testdurchlauf insgesamt noch verbesserungswürdig war. Diese Ergebnisse sind jedoch nicht überraschend, da dies der erste Testlauf war und die Modelle mit einem kontinuierlichen Konversationskontext arbeiteten, der potenziell die Qualität ihrer Antworten beeinflusste.

## **Bert\_Score**:

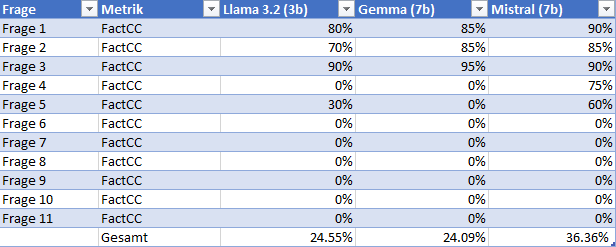
In der Kategorie BERTScore, die die semantische Übereinstimmung der Modellantworten mit den Referenzantworten bewertet, erzielte Llama mit einem Wert von 0.6814 die höchste Punktzahl und ging als Sieger dieser Kategorie hervor. Gemma lag mit einem Wert von 0.6692 im Mittelfeld, während Mistral mit einem Wert von 0.6294 die niedrigste Punktzahl erreichte.

Das Abschneiden von Llama in dieser Kategorie zeigt, dass es im ersten Testdurchlauf die höchste Fähigkeit unter den getesteten Modellen besaß, den semantischen Inhalt der Referenzantworten zu erfassen und wiederzugeben. Ein hoher BERTScore deutet darauf hin, dass Llama im Vergleich zu Gemma und Mistral die Relevanz und den Sinn der ursprünglichen Inhalte am besten verstanden hat, selbst wenn die exakte Wortwahl oder Syntax variierte. Dies spricht für eine solide semantische Analyse- und Interpretationsfähigkeit des Modells.

Jedoch sind die Unterschiede zwischen Llama und Gemma in dieser Kategorie vergleichsweise gering. Dies könnte darauf hindeuten, dass beide Modelle ähnliche Stärken im Umgang mit den semantischen Aspekten der Antworten haben. Das Ergebnis von Mistral hingegen weist darauf hin, dass es Schwierigkeiten hatte, die Bedeutung der Referenzantworten adäquat zu erfassen, was auf Schwächen in der semantischen Generalisierungsfähigkeit hindeuten könnte.

Es ist zudem wichtig anzumerken, dass, obwohl Llama in dieser Kategorie gewonnen hat, die Punktzahlen im Gesamten auf ein moderates Leistungsniveau hindeuten. Dies ist angesichts des ersten Testdurchlaufs und der potenziellen Einschränkungen durch den kontinuierlichen Konversationskontext nicht überraschend. Ein besseres Abschneiden der Modelle könnte in zukünftigen Tests erwartet werden, wenn Maßnahmen zur Vermeidung von Kontextverhaftung umgesetzt werden.

## **FactCC:**

****

Die Ergebnisse der FactCC-Analyse im ersten Testdurchlauf zeigen deutliche Schwächen der getesteten Modelle in der faktischen Konsistenz. Mistral erreichte mit 36.06% den höchsten Durchschnittswert, während Llama und Gemma mit 24.55% bzw. 24.09% nahezu gleichauf lagen. Diese Ergebnisse verdeutlichen, dass keines der Modelle in der Lage war, konstant faktisch korrekte Antworten zu liefern.

Ein genauerer Blick auf die Verteilung der FactCC-Werte zeigt, dass die Modelle in den Fällen, in denen sie die Fragen korrekt verstanden haben, eine hohe faktische Übereinstimmung erzielten, die sich im Bereich von 70% bis 90% bewegte. Dies unterstreicht, dass die Modelle grundsätzlich dazu fähig sind, faktisch konsistente Antworten zu geben, sofern die Frageinhalte richtig interpretiert werden. Allerdings zeigen mehr als die Hälfte der Fragen, die von allen Modellen falsch verstanden wurden, eine faktische Übereinstimmung von 0%. Dies weist auf grundlegende Schwierigkeiten bei der Interpretation der Fragen hin.

Die zwei beobachteten Ausreißer bei den Modellen – hohe FactCC-Werte trotz einer offensichtlich falschen Frageinterpretation – könnten auf zufällige Übereinstimmungen oder fehlerhafte Gewichtung bestimmter Fakten zurückzuführen sein. Solche Abweichungen verdeutlichen die Grenzen der Modelle, komplexe oder kontextabhängige Informationen präzise zu verarbeiten.

Diese Ergebnisse spiegeln nicht nur die Einschränkungen der Modelle im ersten Testdurchlauf wider, sondern heben auch die Bedeutung der Kontextsteuerung und klarer Fragestellungen hervor. Der kontinuierliche Konversationskontext könnte zu einer Fehlfokussierung und Verwirrung bei der Interpretation der Fragen geführt haben, was wiederum die faktische Konsistenz der Antworten negativ beeinflusst hat.

## **Gesamtergebnis:**



Nach der Berechnung der gewichteten Gesamtscores unter Berücksichtigung von ROUGE, BERTScore und FactCC zeigt sich, dass Mistral mit einem Gesamtscore von 0.1434 als klarer Sieger aus dem ersten Testdurchlauf hervorgeht. Dieser Erfolg ist in erster Linie auf die vergleichsweise hohe Leistung in der Kategorie FactCC zurückzuführen, in der Mistral mit 36.36% die höchste faktische Konsistenz erzielte. Obwohl das Modell in den Kategorien ROUGE und BERTScore nicht die besten Ergebnisse lieferte, konnte es sich durch die stärkere Gewichtung der FactCC-Metrik insgesamt durchsetzen. Dies verdeutlicht die zentrale Rolle der Faktentreue in der Bewertung.

Llama erreichte mit einem Gesamtscore von 0.0908 den zweiten Platz, knapp vor Gemma mit 0.0872. Beide Modelle zeigten ähnliche Leistungen in der Kategorie FactCC (24.55% bzw. 24.09%) und lagen auch in ROUGE und BERTScore auf einem ähnlichen Niveau. Dennoch konnte Llama sich durch geringfügig bessere Werte in den beiden sprach- und semantikbezogenen Kategorien einen leichten Vorsprung verschaffen.

Die vergleichsweise niedrigen Gesamtscores aller Modelle verdeutlichen jedoch, dass sie im ersten Testdurchlauf Schwierigkeiten hatten, die gestellten Fragen konsistent und korrekt zu beantworten. Dies wird besonders durch die stark schwankenden FactCC-Werte deutlich, die zwischen 0% und 90% variierten. Während einige Fragen von den Modellen gut verstanden und beantwortet wurden, führte eine fehlerhafte Interpretation der meisten Fragen zu erheblichen Einbußen in der faktischen Konsistenz. Die Ergebnisse machen deutlich, dass die Modelle zwar in der Lage sind, bei klaren Fragestellungen hohe faktische Übereinstimmungen zu erzielen, dass jedoch ihre Fähigkeit, Kontextinformationen richtig zu interpretieren, im ersten Testdurchlauf stark eingeschränkt war.

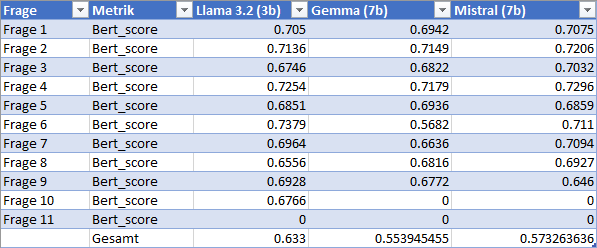
Die Ergebnisse dieses Testdurchlaufs liefern wichtige Erkenntnisse für die weitere Evaluierung. Die Schwächen, die durch den kontinuierlichen Konversationskontext hervorgerufen wurden, könnten im zweiten Testlauf, bei dem der Kontext vor jeder neuen Frage zurückgesetzt wurde, reduziert werden. Dies könnte zu einer Verbesserung der faktischen Konsistenz und der Gesamtleistung der Modelle führen.

## **Ergebisse Testlauf 2:**

## **Rouge:**

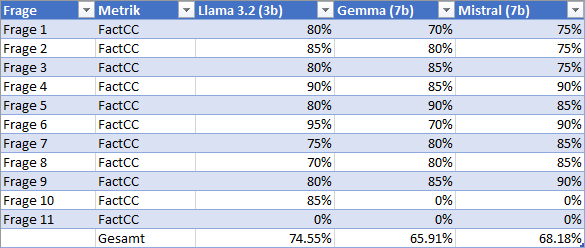
In der Kategorie ROUGE zeigte Llama die beste Leistung mit einem Gesamtscore von 0.2002, während Gemma mit 0.1644 den niedrigsten Wert erzielte. Mistral lag mit 0.1791 im Mittelfeld. Die Ergebnisse zeigen, dass Llama im zweiten Testlauf die sprachliche Präzision und Kohärenz der Antworten am besten beibehalten konnte. Dennoch bleiben die Werte in dieser Kategorie insgesamt eher niedrig, was auf sprachliche und syntaktische Herausforderungen bei allen Modellen hinweist.

## **Bert\_Score**:



Auch in der Kategorie BERTScore erzielte Llama mit 0.633 die höchste Punktzahl, was auf eine verbesserte semantische Übereinstimmung mit den Referenzantworten hindeutet. Mistral und Gemma blieben mit Werten von 0.5733 bzw. 0.5540 hinter Llama zurück. Diese Ergebnisse spiegeln wider, dass Llama im zweiten Testlauf am besten in der Lage war, die Bedeutung der Fragen und Referenzantworten konsistent zu erfassen.

## **FactCC:**



Bei der Bewertung der faktischen Konsistenz zeigte Llama ebenfalls die stärkste Leistung mit einem Durchschnittswert von 74.55%. Mistral folgte mit 68.18%, während Gemma mit 65.91% die niedrigste Punktzahl erreichte. Die FactCC-Werte im zweiten Testlauf sind insgesamt deutlich höher als im ersten Testlauf, was darauf hindeutet, dass der zurückgesetzte Kontext den Modellen geholfen hat, die Fragen präziser und korrekter zu beantworten. Frage 10 und 11 bezogen sich nicht auf einen spezifischen Spieler sondern jeweils auf ein Team. Obwohl in der Vektordatenbank Informationen über diese vorhanden waren konnten die Modelle die Fragen zu diesen fast gar nicht beantworten.

## **Gesamtergebnis:**

Llama erzielte mit einem gewichteten Gesamtscore von 0.3923 den höchsten Wert und konnte sich damit als Sieger des zweiten Testlaufs etablieren. Mistral folgte mit 0.3259, während Gemma mit 0.3026 den dritten Platz belegte. Diese Ergebnisse verdeutlichen, dass die Modelle im zweiten Testlauf, durch die Vermeidung von Kontextverhaftung, insbesondere bei der faktischen Konsistenz bessere Leistungen erzielen konnten. Die stärkere Gewichtung der FactCC-Metrik spielte hierbei eine zentrale Rolle, da faktische Richtigkeit ein wesentlicher Aspekt der Bewertung war.

Die Verbesserungen in den FactCC-Werten und die relativ höheren BERTScores zeigen, dass die Kontextrücksetzung ein effektives Mittel war, um die Modelle auf die jeweilige Frage zu fokussieren und damit ihre Leistung in den einzelnen Kategorien zu steigern. Dennoch gibt es weiterhin Potenzial zur Verbesserung, insbesondere im Bereich der sprachlichen Präzision und der Konsistenz über alle Kategorien hinweg.

Darüber hinaus wurde die durchschnittliche Antwortzeit der verschiedenen Modelle gemessen. Hier zeigte Mistral mit einer Antwortzeit von etwa 5 Sekunden die schnellste Reaktionszeit, gefolgt von Llama mit 8,5 Sekunden. Gemma lag mit einer durchschnittlichen Antwortzeit von 15,3 Sekunden pro Prompt deutlich hinter den beiden anderen Modellen zurück.

## **Fazit:**

Die Ergebnisse aus beiden Testdurchläufen zeigen, dass Llama und Mistral in den verschiedenen Kategorien jeweils ihre Stärken haben, während Gemma insgesamt etwas hinterherhinkt. Der zweite Testlauf verdeutlicht jedoch, dass die Leistung aller Modelle durch eine klare Kontextrücksetzung verbessert werden kann, wobei weiterhin Optimierungen, insbesondere in der sprachlichen Präzision und Konsistenz, erforderlich sind. Mistral überzeugt zusätzlich durch seine schnelle Antwortzeit, auch wenn dies in der aktuellen Anwendung nicht von entscheidender Bedeutung ist.