

LLM Performance Evaluation Report

Team 25

December 16, 2025

1 Case-Based Model Evaluation

This section presents a detailed, case-based evaluation of three large language models (LLaMA 3.3, Gemma 3, and MistralAI) using knowledge-graph (KG) grounded question answering. Each case includes the input question, full model-generated answers, and a justified qualitative and quantitative analysis.

1.1 Evaluation Criteria

Models were evaluated using:

- **Quantitative metrics:** accuracy, response time, token usage, and cost
- **Qualitative metrics:** relevance, correctness, completeness, naturalness, and confidence

Qualitative scores were assigned via human evaluation on a 1–5 Likert scale.

1.2 Quantitative Performance Comparison

Table 1: Quantitative Metrics Across Models

Model	Accuracy	Response Time (s)	Tokens	Cost
LLaMA 3.3	0.77	6.35	2023	Free
Gemma 3	0.63	5.01	2568	Free
MistralAI	0.85	4.65	2853	Free

Table 2: Average Qualitative Scores Across Test Cases

Model	Relevance	Correctness	Completeness	Naturalness	Confidence
LLaMA 3.3	4.5	4.0	3.0	4.0	3.5
Gemma 3	4.5	3.5	3.5	4.5	3.5
MistralAI	5.0	5.0	4.0	4.0	4.5

1.3 Test Case 1: Identifying Top Midfielders by Assists

1.3.1 Question

Who are the top midfielders from Arsenal in the 2022–23 season with the most assists?

1.3.2 LLaMA 3.3 Response

Based on the context, the top midfielders from Arsenal in 2022-23 with the most assists are:

1. Bukayo Saka with 12 assists
2. Martin Ødegaard is not listed with assists in the context, but Gabriel Martinelli Silva is not listed with assists either, however, another player from Arsenal is listed
3. No other Arsenal midfielders are listed with assists in the context.

Analysis Correctly identifies Bukayo Saka as a top contributor. Introduces unnecessary uncertainty and references players without assist data, reducing completeness. The answer is relevant and factual for the named player but verbose and hedged.

Justification

- Relevance: Very high
- Correctness: High (Saka correct; others hedged)
- Completeness: Moderate (only Saka verified)
- Naturalness: Good
- Confidence: Moderate (hedging lowers confidence)

1.3.3 Gemma 3 Response

Based on the provided context, here are the top midfielders from Arsenal with the most assists in 2022-23: - Bukayo Saka: 12 assists, creativity score of 322.8 - Leandro Trossard: 13 assists, creativity score of 211.1

Analysis Concise and readable, but incorrectly lists Leandro Trossard (not an Arsenal midfielder) and does not clarify that stats may be season-wide, not Arsenal-specific.

Justification

- Relevance: High
- Correctness: Moderate (role/stat ambiguity)
- Completeness: Moderate (misassigned player)
- Naturalness: Very high
- Confidence: Moderate

1.3.4 MistralAI Response

Based on the context provided, the top midfielders from Arsenal in the 2022-23 season with the most assists are: 1. Bukayo Saka - 12 assists 2. Martin Ødegaard - Not listed in the top playmakers, but he is one of the top midfielders by total points. 3. Gabriel Martinelli Silva - Not listed in the top playmakers, but he is one of the top midfielders by total points.

The only Arsenal midfielder explicitly mentioned in the top playmakers list is Bukayo Saka with 12 assists. The other Arsenal midfielders listed (Martin Ødegaard and Gabriel Martinelli Silva) are not mentioned in the top playmakers section, so their assist numbers are not provided in the context.

Analysis Strong KG-awareness; clearly separates verified facts from unverified. Avoids hallucination and explains context limitations, though longer than necessary.

Justification

- Relevance: Very high
- Correctness: Very high (no hallucinations)
- Completeness: Moderate
- Naturalness: Good
- Confidence: High

1.4 Test Case 2: Team-Specific Player Performance

1.4.1 Question

How many goals and assists did Leandro Trossard record specifically for Arsenal in the 2022–23 Premier League season, and in which match did he register a hat-trick of assists?

1.4.2 LLaMA 3.3 Response

I'm afraid I don't have that specific information. The provided context does not contain data on Leandro Trossard's performance for Arsenal in the 2022-23 Premier League season, nor does it mention a specific match where he registered a 'hat-trick' of assists. The context only provides general statistics for Leandro Trossard, but it does not differentiate between his performances for Brighton and Arsenal.

Analysis Correctly recognizes that the KG lacks team-specific breakdown. Avoids hallucination and maintains factual accuracy, though verbose.

Justification

- Relevance: Very high
- Correctness: Very high
- Completeness: Moderate
- Naturalness: Good
- Confidence: High

1.4.3 Gemma 3 Response

The provided context only contains information about Leandro Trossard's overall stats for the 2022-23 season, not a breakdown by team. It states he recorded 8 goals and 13 assists. The context does not mention any specific match where he registered a "hat-trick" of assists.

Analysis Provides overall stats but not team-specific. Potentially misleading for Arsenal-only question, reducing correctness.

Justification

- Relevance: Moderate
- Correctness: Low to moderate
- Completeness: Moderate
- Naturalness: Good
- Confidence: Low

1.4.4 MistralAI Response

The provided context does not include specific details about Leandro Trossard’s performance for Arsenal separately from his overall season statistics. Therefore, I cannot provide the exact number of goals and assists he recorded specifically for Arsenal in the 2022-23 Premier League season or identify the match in which he registered a ‘hat-trick’ of assists. The context only provides his aggregated statistics for the entire season, including both his time at Brighton and Arsenal.

Analysis Explicitly states KG limitations, avoids assumptions. Factually precise, clearly communicates uncertainty.

Justification

- Relevance: Very high
- Correctness: Very high
- Completeness: Moderate
- Naturalness: Good
- Confidence: Very high

1.5 Strengths and Weaknesses Summary

Table 3: Model Strengths and Weaknesses

Model	Strengths	Weaknesses
LLaMA 3.3	Token efficiency, reasonable correctness	Limited completeness, verbose hedging
Gemma 3	Fluent, natural language	Factual inaccuracies, misleading assumptions
MistralAI	High accuracy, KG-awareness, avoids hallucinations	Slightly verbose, longer explanations

1.6 Overall Discussion

Quantitatively, MistralAI achieved the highest accuracy and lowest response time. Qualitatively, it consistently avoided hallucinations and demonstrated strong awareness of KG limitations. Gemma 3 excelled in fluency but made factual assumptions. LLaMA 3.3 balanced efficiency and correctness but lacked completeness.

1.7 Conclusion

For knowledge-graph grounded question answering, MistralAI is the most reliable model, especially when handling incomplete or ambiguous data.