LLM Performance Evaluation Report

Team 25

December 16, 2025

1 Model Evaluation and Comparative Analysis

This section presents a comprehensive evaluation of three large language models (LLaMA 3.3, Gemma 3, and MistralAI) for knowledge-graph (KG) grounded question answering. The evaluation combines quantitative performance metrics with qualitative human analysis to justify model behavior and suitability for the target use case.

1.1 Evaluation Objectives

The primary objective is to assess how effectively each model:

- Retrieves and reasons over KG-grounded information
- Avoids hallucination when required facts are missing
- Produces clear, relevant, and natural language responses

1.2 Experimental Setup

Two test cases were designed using football statistics grounded in a structured KG. Each question requires either precise entity filtering or recognition of missing information in the KG, making them suitable for evaluating factual robustness.

1.3 Quantitative Evaluation Metrics

The following quantitative metrics were collected:

- **Accuracy**: correctness of the final answer
- **Response Time**: latency in seconds
- **Token Usage**: sum of input and output tokens
- **Cost**: monetary cost of inference

Table 1: Quantitative Performance Comparison

Model	Accuracy	Resp. 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

Quantitative Analysis MistralAI achieves the highest overall accuracy while also exhibiting the lowest average response time. LLaMA 3.3 demonstrates strong token efficiency, whereas Gemma 3 consumes the highest number of tokens with comparatively lower accuracy, suggesting inefficiencies in reasoning over the KG context.

1.4 Qualitative Evaluation Criteria

Human evaluators scored model responses on a 1–5 Likert scale using the following dimensions:

- Relevance
- Correctness
- Completeness
- Naturalness
- Confidence

Table 2: Average Qualitative Scores Across Test Cases

Model	Rel.	Corr.	Comp.	Nat.	Conf.
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**

Qualitative Analysis While Gemma 3 excels in linguistic fluency, it occasionally introduces ambiguous or unverified facts. LLaMA 3.3 maintains reasonable balance but lacks depth in multi-constraint reasoning. MistralAI consistently demonstrates factual discipline, explicit acknowledgment of missing KG data, and high confidence without hallucination.

1.5 Case-Based Qualitative Justification

1.5.1 Test Case 1: Top Arsenal Midfielders by Assists

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

Discussion LLaMA 3.3 correctly identifies Bukayo Saka but introduces uncertainty by mentioning players without assist statistics. Gemma 3 provides a concise list but incorrectly assumes positional roles. MistralAI limits its answer strictly to verifiable KG facts, demonstrating superior KG-awareness despite reduced verbosity.

1.5.2 Test Case 2: Team-Specific Player Performance

Question: *How many goals and assists did Leandro Trossard record specifically for Arsenal, and in which match did he register a hat-trick of assists?*

Discussion Both LLaMA 3.3 and MistralAI correctly state that the KG lacks team-specific breakdowns. Gemma 3 introduces overall statistics, which may mislead users. This behavior highlights the importance of explicit uncertainty handling in KG-based systems.

1.6 Overall Model Comparison

Table 3: Strengths and Weaknesses Summary

Model	Strengths	Weaknesses
LLaMA 3.3	Token efficiency, reasonable correctness	Limited completeness, mild hedging
Gemma 3	Natural language fluency	Occasional factual ambiguity
MistralAI	High accuracy, KG-awareness, low latency	Slight verbosity

1.7 Conclusion

The combined quantitative and qualitative analysis demonstrates that MistralAI is the most suitable model for KG-grounded question answering in this setting. Its ability to maintain factual correctness, acknowledge missing information, and respond efficiently makes it the preferred choice for deployment.