Qualitative Evaluation: Human Assessment of LLM Responses

1. meta-llama/Llama-3.1-8B-Instruct

Answer Quality:
This model produced clear, well-structured, and informative responses, particularly for hotel recommendation and filtering queries. When sufficient context was provided, answers were complete and useful. In cases of missing context, the model appropriately limited its output.

Relevance:
Responses were highly aligned with the user's intent, consistently addressing the core requirements of each query without introducing irrelevant information.

Naturalness:
The language was fluent, coherent, and professional, closely resembling a human-written travel assistant response.

Correctness:
The model demonstrated strong factual correctness when context information was available and avoided hallucinations when data was insufficient.

---

2. mistralai/Mistral-7B-Instruct-v0.2

Answer Quality:
Mistral provided balanced and concise answers, offering enough detail to be helpful while remaining succinct. It performed well on structured recommendation tasks.

Relevance:
Responses were consistently relevant and stayed focused on the question being asked.

Naturalness:
The model exhibited natural sentence flow and clear phrasing, though slightly less expressive than LLaMA in some cases.

Correctness:
Correctness was generally high, with the model appropriately relying on the provided context and avoiding unsupported claims.

---

3. deepseek-ai/DeepSeek-V3.2

Answer Quality:
DeepSeek generated moderately detailed responses that were informative but sometimes generic. While answers were useful, they occasionally lacked specificity compared to larger models.

Relevance:
The model maintained good topical relevance, directly responding to the user queries without significant deviation.

Naturalness:
Language quality was acceptable but slightly more formal and less conversational, making responses feel more system-like.

Correctness:
DeepSeek showed strong adherence to the provided context, consistently avoiding hallucinations and unsupported information.

---

4. openai/gpt-oss-20b

Answer Quality:
This model produced high-quality, comprehensive responses, often providing the most complete and insightful answers among the evaluated models.

Relevance:
Responses were highly relevant and well-aligned with the user's intent, even for more complex or multi-part queries.

Naturalness:
The language was highly natural, coherent, and conversational, closely matching human-like explanations.

Correctness:
The model demonstrated excellent factual accuracy, effectively integrating context information while avoiding unsupported assumptions.