Gen-Ai

LLM Model Comparison Report

Submitted By

Khuram Iqbal 2022-CS-48

Program Supervisor

Xavor Corporation

Department of Computer Science

University of Engineering and Technology Lahore, Pakistan

1 Summary

Zero-Shot: Both models provided accurate sentiment classification but failed to correctly solve the budget allocation problem.

Few-Shot: Both models correctly implemented the factorial function, demonstrating their capability to learn from examples.

Chain-of-Thought: Both models struggled with the odd numbers summation task, but HuggingfaceH4 Zephyr-7b-beta correctly solved the apples calculation problem, whereas LLaMa v2 did not.

2 Model Responses

1. **Zero-Shot - Sentiment Classification**
   - HuggingfaceH4 Zephyr-7b-beta Response: Neutral. The use of "okay" indicates a lack of strong positive or negative feelings.
   - LLaMa v2 Response: Neutral, with additional explanations leaning slightly towards positive and negative.
2. **Zero-Shot - Budget Allocation**
   - HuggingfaceH4 Zephyr-7b-beta Response: Provided equations and solution but with errors in the calculations and verification steps.
   - LLaMa v2 Response: Provided equations and solution but with calculation errors and verification issues.
3. **Few-Shot - Factorial Function**
   - HuggingfaceH4 Zephyr-7b-beta Response: Provided a correct recursive Python function.
   - LLaMa v2 Response: Provided a correct recursive Python function.
4. **Chain-of-Thought - Odd Numbers Summation**
   - HuggingfaceH4 Zephyr-7b-beta Response: Incorrect explanation, adding both odd and even numbers together.
   - LLaMa v2 Response: Incorrectly adds both odd and even numbers together.
5. **Chain-of-Thought - Apples Calculation**
   - HuggingfaceH4 Zephyr-7b-beta Response: Correctly calculated the remaining apples.
   - LLaMa v2 Response: Incorrectly calculates the remaining apples.

3 Conclusion

HuggingfaceH4 Zephyr-7b-beta: Generally performed better in Chain-of-Thought prompts, while both models excelled equally in Few-Shot prompts.

LLaMa v2: Both models had issues with Zero-Shot prompts, especially with more complex tasks like the budget allocation problem. For sentiment classification, both models performed adequately.