Comparative Study of Various Large Language Models (LLMs) in the Task of Answering Questions Based on Tabular Image Data

Introduction

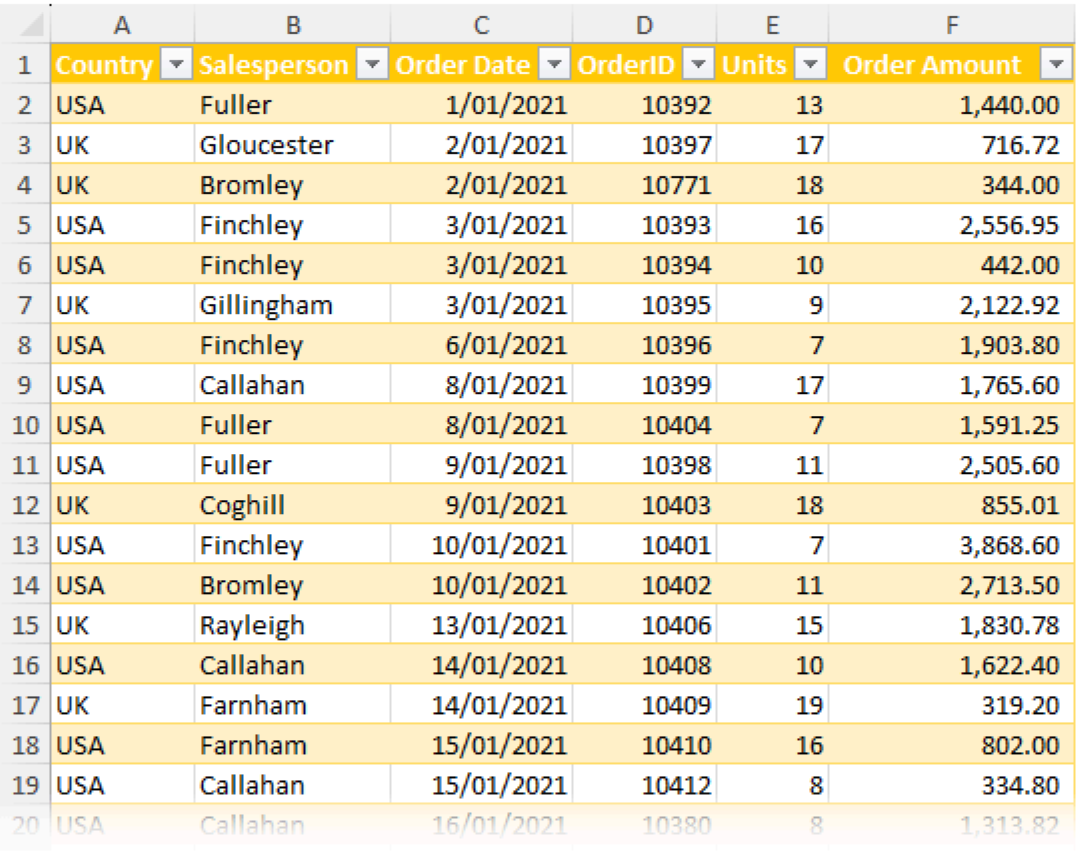
The task involves analyzing tabular data from an image to determine the salesperson with the highest total order amount. This report compares the performance of three different Large Language Models (LLMs) in achieving this task: OpenAI's GPT-4o-mini, Llama 3.1 8B, and Mistral 7B. It considers the different computational resources available for each model, noting that GPT-4o-mini runs on high-end servers, while Llama 3.1 8B and Mistral 7B operate on a local machine with limited computational power. The text data was extracted from the image using **PyTesseract**, an optical character recognition (OCR) tool, before being fed into the LLMs for analysis.

Text Extraction Using PyTesseract

Approach:

1. **Image Processing:**
    - The image containing the tabular data was processed using PyTesseract, an OCR library in Python that converts images to text.
    - The extracted text was then given for further analysis by the LLMs.
2. **Feeding Data to LLMs:**
    - The extracted data was provided as input to each of the LLMs to perform the analysis and compute the total order amounts for each salesperson.
    - For **Mistral 7B** and **Llama 3.1 8B**, **Ollama** was used, which allows these models to run on a local machine. This involved installing Ollama and then downloading the Mistral 7B and Llama 3.1 models to facilitate their operation in a constrained computational environment.

Input and Prompt



	A	B	C	D	E	F
1	Country	Salesperson	Order Date	OrderID	Units	Order Amount
2	USA	Fuller	1/01/2021	10392	13	1,440.00
3	UK	Gloucester	2/01/2021	10397	17	716.72
4	UK	Bromley	2/01/2021	10771	18	344.00
5	USA	Finchley	3/01/2021	10393	16	2,556.95
6	USA	Finchley	3/01/2021	10394	10	442.00
7	UK	Gillingham	3/01/2021	10395	9	2,122.92
8	USA	Finchley	6/01/2021	10396	7	1,903.80
9	USA	Callahan	8/01/2021	10399	17	1,765.60
10	USA	Fuller	8/01/2021	10404	7	1,591.25
11	USA	Fuller	9/01/2021	10398	11	2,505.60
12	UK	Coghill	9/01/2021	10403	18	855.01
13	USA	Finchley	10/01/2021	10401	7	3,868.60
14	USA	Bromley	10/01/2021	10402	11	2,713.50
15	UK	Rayleigh	13/01/2021	10406	15	1,830.78
16	USA	Callahan	14/01/2021	10408	10	1,622.40
17	UK	Farnham	14/01/2021	10409	19	319.20
18	USA	Farnham	15/01/2021	10410	16	802.00
19	USA	Callahan	15/01/2021	10412	8	334.80
20	USA	Callahan	16/01/2021	10380	8	1,313.82

Question: Can you tell me which salesperson made highest order amount. You need to add all order amount and tell me who made the highest order amount?

Prompt: ;
[
{"role": "system", "content": "You are a helpful AI assistant. You will be given an extracted tabular data. You are tasked to answer the user questions"},
{"role": "user", "content": f"Extracted text: {extracted_text}\n Question: {question}"}
]

Models Evaluated

1. **OpenAI GPT-4o-mini**
   - **Computational Environment:** High-end servers with substantial processing capabilities and memory.
   - **Task Performance:**
     - Correctly identified the salesperson with the highest order amount.

        - Provided a detailed breakdown of the order amounts for each salesperson.
        - Achieved high accuracy in calculations and clear presentation of results.

2. **Llama 3.1 8B**
    - **Computational Environment:** Local machine with Intel i5 processor and 8GB RAM.
    - **Task Performance:**
        - Incorrect calculation of the total order amounts.
        - Misidentification of the highest total due to summing all salesperson totals together.
        - Provided a breakdown of each salesperson's order amount but with errors.
3. **Mistral 7B**
    - **Computational Environment:** Local machine with Intel i5 processor and 8GB RAM.
    - **Task Performance:**
        - Incorrectly identified the total order amount for the top salesperson.
        - Demonstrated a logical approach with structured coding methodology.
        - Encountered calculation errors due to misinterpretation of data.

Comparative Analysis

Accuracy and Correctness

- **OpenAI GPT-4o-mini:**
    - Achieved the highest accuracy with correct identification of the top salesperson, Finchley, with a total order amount of **$8,771.35**.
    - Minimal errors in data interpretation and computation.
- **Llama 3.1 8B:**
    - Failed to produce the correct total due to aggregation errors.
    - Incorrectly presented Finchley with a total of **$21,066.39**, which was a result of summing all salespersons' totals rather than Finchley's alone.
- **Mistral 7B:**
    - Miscalculated the total order amount, identifying Finchley with **$10,675.35**.
    - Showed a strong methodological approach but failed in execution and calculation accuracy.

Methodology and Approach

- **OpenAI GPT-4o-mini:**
    - Used a clear, detailed methodology that accurately parsed and calculated totals.
    - Provided detailed breakdowns that helped verify each step of the process.
- **Llama 3.1 8B:**

    - Attempted a straightforward aggregation approach but suffered from missteps in summation logic.
    - Errors in handling the data led to incorrect outcomes.
- **Mistral 7B:**
    - Implemented a structured code-driven approach that was methodologically sound.
    - Suffered from misinterpretation and execution errors that skewed results.

Performance in Context of Computational Resources

- **OpenAI GPT-4o-mini:**
    - Utilized high computational resources to deliver precise and detailed outputs.
    - Benefited from greater processing power, which aided in handling complex tasks more effectively.
- **Llama 3.1 8B and Mistral 7B:**
    - Operated under significantly constrained computational environments, affecting processing speed and memory management.
    - Performance was hampered by hardware limitations, potentially contributing to the errors observed.
    - Demonstrated that even under constraints, logical approaches can be formulated, though execution suffered.

Conclusions and Recommendations

Overall, this comparative study highlights the impact of computational resources on LLM performance and suggests pathways for enhancing efficiency in models operating under resource constraints. Proper pre-processing and data extraction play a crucial role in ensuring accurate results across different computational environments.