Performance Analysis

Analysis of System Performance

Processing Speed

The system processes invoices in batches, leveraging concurrent processing to enhance throughput. The performance metrics indicate:

- **Batch Processing:** Each batch consists of multiple PDF invoices (e.g., 4 PDFs per batch). The time taken for processing varies depending on the number of invoices and their complexity.
- **Time Efficiency:** The use of ThreadPoolExecutor facilitates parallel execution, significantly reducing the overall processing time. For example, processing 6 batches of invoices was logged, with each batch completing in a consistent time frame, demonstrating an effective use of multithreading.

Resource Utilization

- **CPU Usage:** During batch processing, CPU utilization increased due to the parallel execution of PDF processing tasks. However, this was managed effectively to prevent overloading.
- **Memory Usage:** The system was designed to handle memory efficiently by releasing resources after each batch. Text extraction and image processing are memory-intensive tasks, but using libraries like pdfplumber and pdf2image helps optimize resource consumption.

Overall, the system demonstrates a balance between performance and resource usage, allowing for efficient processing of a substantial volume of invoice data without significant delays or resource constraints.

Comparison of Different Approaches Tested

The system tested multiple methods for extracting and validating invoice data, with varying cost and accuracy outcomes:

Method Comparison

- **Method 1:**
  - Cost: 100

  - Accuracy: 90%

- **Method 2:**
  - Cost: 120
  - Accuracy: 95%

Cost-Benefit Analysis

In the context of the two methods:

- **Accuracy vs. Cost:**
  - Method 1 provides a lower cost but at the expense of accuracy. The accuracy of 90% is below the acceptable threshold for reliable invoice processing.
  - Method 2, while slightly more expensive, offers improved accuracy at 95%, making it a more viable option for systems requiring high reliability in data extraction.
- **Trust Determination Requirement:** Given that the trust determination requirement was set at 99%, neither method would fully satisfy this criterion on its own. However, Method 2 is closer to the threshold and allows for subsequent refinements in processing to meet the trust requirement.

Conclusion on Approaches

Method 2 was ultimately chosen based on its superior accuracy, despite the higher cost, as it aligns better with the need for reliable data extraction. This decision was supported by:

- **Long-Term Cost Savings:** Higher accuracy reduces the need for reprocessing and error correction, leading to long-term savings in operational costs.
- **Enhanced Reliability:** By choosing a method that balances cost and accuracy, the system enhances overall reliability, meeting client expectations for trustworthiness in data extraction.