Invoice Reimbursement Analysis

LLM Selection: Google Gemini Pro

I chose Google Gemini's LLM for this project because:
- Strong text analysis capabilities
- Reliable JSON output formatting
- Good policy comprehension
- Cost-effective for high-volume processing
- Fast response times with `gemini-1.5-flash` model

File Processing Approach

PDF Processing
- Using `pdfplumber` library for text extraction
- Handles different PDF formats effectively
- Extracts structured text maintaining formatting
- Error handling for scanned/ image-based PDFs

```python
def extract_text_from_pdf(pdf_content: bytes) -> str:
    try:
        text = ""
        with pdfplumber.open(io.BytesIO(pdf_content)) as pdf:
            for page in pdf.pages:
                text += page.extract_text() or ""
        return text.strip()
    except Exception as e:
        logger.error(f"PDF extraction error: {str(e)}")
        return ""
```

ZIP File Handling
- Asynchronous processing using FastAPI
- Extracts all PDFs from ZIP archives
- Maintains folder structure information
- Validates file types and content

Prompt Optimization

Original vs Optimized Prompt

1. **Structured Analysis Process**
   a. Added numbered steps for systematic review
   b. Clear breakdown of each analysis phase
   c. Ensures consistent evaluation across invoices
2. **Response Format Standardization**
   a. Defined strict JSON structure
   b. Required fields with clear data types
   c. Consistent status values
3. **Policy Compliance Focus**
   a. Explicit reminder about pre-approval status
   b. Required citation of policy clauses
   c. Clear decision criteria
4. **Calculation Rules**
   a. Integer-only amounts
   b. Double validation requirement
   c. Clear partial reimbursement rules

Key Improvements
1. **Process Clarity**
```
ANALYSIS PROCESS:
1. Invoice Details Analysis
2. Policy Compliance Check
3. Reimbursement Calculation
4. Final Determination
```

2. **Decision Framework**
```
Choose one status:
- "Fully Reimbursed": Meets all policy requirements
- "Partially Reimbursed": Eligible but exceeds limits
- "Declined": Violates policy restrictions
```

3. **Validation Requirements**
```
IMPORTANT RULES:
- Follow policy rules exactly
- Always cite specific policy clauses
- Double-check all calculations
- Use only integer amounts
```

The optimized prompt ensures:
● Consistent analysis methodology
● Clear decision-making process
● Accurate JSON output format
● Traceable reasoning through policy citations