**Invoice data extraction with LlamaParse and**

OpenAI (1)

**Workflow: Intelligent Invoice Data Extraction & Reconciliation**

Overview

This workflow automates the process of extracting structured data from invoice PDFs received via email. It uses AI and web scraping techniques to:

* Download invoices from Gmail.
* Parse the invoice PDFs using LlamaParse.
* Extract key data fields using a large language model (LLM).
* Update a Google Sheet (or similar datastore) with the extracted invoice details.
* Optionally, add labels to processed emails to avoid duplicate processing.

Key Components and Data Flow

1. **Email Trigger and Data Ingestion**
   * **Gmail Trigger ("Receiving Invoices"):**
     + **Function:** Monitors a Gmail inbox for incoming invoice emails with PDF attachments.
     + **Filters:** Only processes emails that have an attachment (typically an invoice) and do not have the label “invoice synced.”
2. **PDF Download and Extraction**
   * **HTTP Request / Download Node:**
     + **Function:** Downloads the attached invoice PDF from Gmail.
   * **Extract from File Node:**
     + **Function:** Parses the downloaded PDF to extract its textual content.
     + **Operation:** Uses PDF processing to convert the file content into text for further analysis.
3. **Advanced PDF Processing with LlamaParse**
   * **Upload to LlamaParse Node:**
     + **Function:** Uploads the PDF file to LlamaParse, a cloud-based PDF parsing service.
     + **Operation:** Sends the PDF as multipart/form-data to LlamaParse's API for advanced parsing, which can handle complex PDFs (tables, figures, etc.).
   * **Get Processing Status Node:**
     + **Function:** Checks the processing status of the uploaded PDF to ensure it has been parsed successfully.
4. **Data Extraction and Structuring**
   * **Apply Data Extraction Rules (LLM Chain):**
     + **Function:** Uses an AI agent to analyze the markdown output from LlamaParse.
     + **Objective:** Extract specific invoice details (invoice date, invoice number, supplier details, line items, prices, etc.) using a guided prompt.
   * **Structured Output Parser Node:**
     + **Function:** Formats the AI's output into a structured JSON format that includes all required fields.
     + **Schema:** Defines fields such as “Invoice date,” “Supplier name,” “Line items,” etc.
5. **Data Aggregation and Update**
   * **Map Output Node:**
     + **Function:** Collects the structured data from the AI extraction process.
   * **Append to Reconciliation Sheet Node:**
     + **Function:** Updates a Google Sheet with the extracted invoice data.
     + **Integration:** Uses the Google Sheets node to append a new row with data like invoice number, supplier name, totals, etc.
   * **Add “invoice synced” Label:**
     + **Function:** Applies a label to the processed email in Gmail to prevent reprocessing.
6. **User Interaction and Feedback**
   * **Respond to Webhook/Feedback:**
     + **Function:** Provides a structured response back to the user or system indicating the extraction status and any next steps.
   * **Sticky Notes (Documentation):**
     + **Function:** Various sticky note nodes are included throughout the workflow to document key steps, provide configuration tips, and offer guidance on customization (e.g., updating API keys, changing filter criteria).

Execution Sequence

1. **Trigger:**  
   The workflow is triggered when a new invoice email is received.
2. **Download & Extract:**  
   The invoice PDF is downloaded and its content is extracted into text.
3. **Advanced Parsing:**  
   The PDF is uploaded to LlamaParse, and once processed, its markdown output is retrieved.
4. **Data Extraction:**  
   The markdown output is fed into an AI agent, which extracts structured invoice data based on predefined rules.
5. **Update Datastore:**  
   Extracted data is mapped and appended to a Google Sheet for reconciliation.
6. **Labeling:**  
   The original invoice email is labeled as “invoice synced” to avoid duplicate processing.
7. **Response:**  
   The workflow may send back a confirmation or log the results for further action.

Benefits

* **Automated Data Processing:**  
  Minimizes manual entry and errors by automatically extracting invoice data from PDFs.
* **Advanced Parsing Capabilities:**  
  Leverages LlamaParse to handle complex PDF structures, ensuring high-quality data extraction.
* **Structured Output:**  
  Uses an AI agent to extract and format key information into a standardized JSON structure.
* **Seamless Integration:**  
  Updates a Google Sheet with extracted data, streamlining financial reconciliation.
* **Efficiency:**  
  Applies labels to processed emails, preventing duplicate work and ensuring efficient workflow execution.