GEN AI Based Email Classification

Revolutionizing Loan Servicing Email Triage

This presentation introduces an innovative approach to loan servicing email triage through the use of Generative AI. Our goal is to showcase how this technology can enhance efficiency, accuracy, and customer satisfaction within the banking industry.

AI TITANS

The Problem: Manual Email Triage Bottleneck

Commercial Bank Lending Services are inundated with servicing requests, receiving thousands of emails each day

Manual review is required for each email to interpret content, identify request type and sub-type, and extract key attributes. This is time-consuming and costly.

🕐 **Time-Consuming**
Manual review is slow.

🐛 **Error-Prone**
Human error is common.

💵 **Costly**

Current Manual Process: A Deep Dive

Servicing agents typically dedicate several minutes to reviewing each email, which contributes to the growing backlog.

This results in delayed response times, increased customer dissatisfaction, and a higher likelihood of routing errors.

The current system faces challenges such as backlogs, errors, and compliance risks, leading to inefficiencies and potential lost productivity.

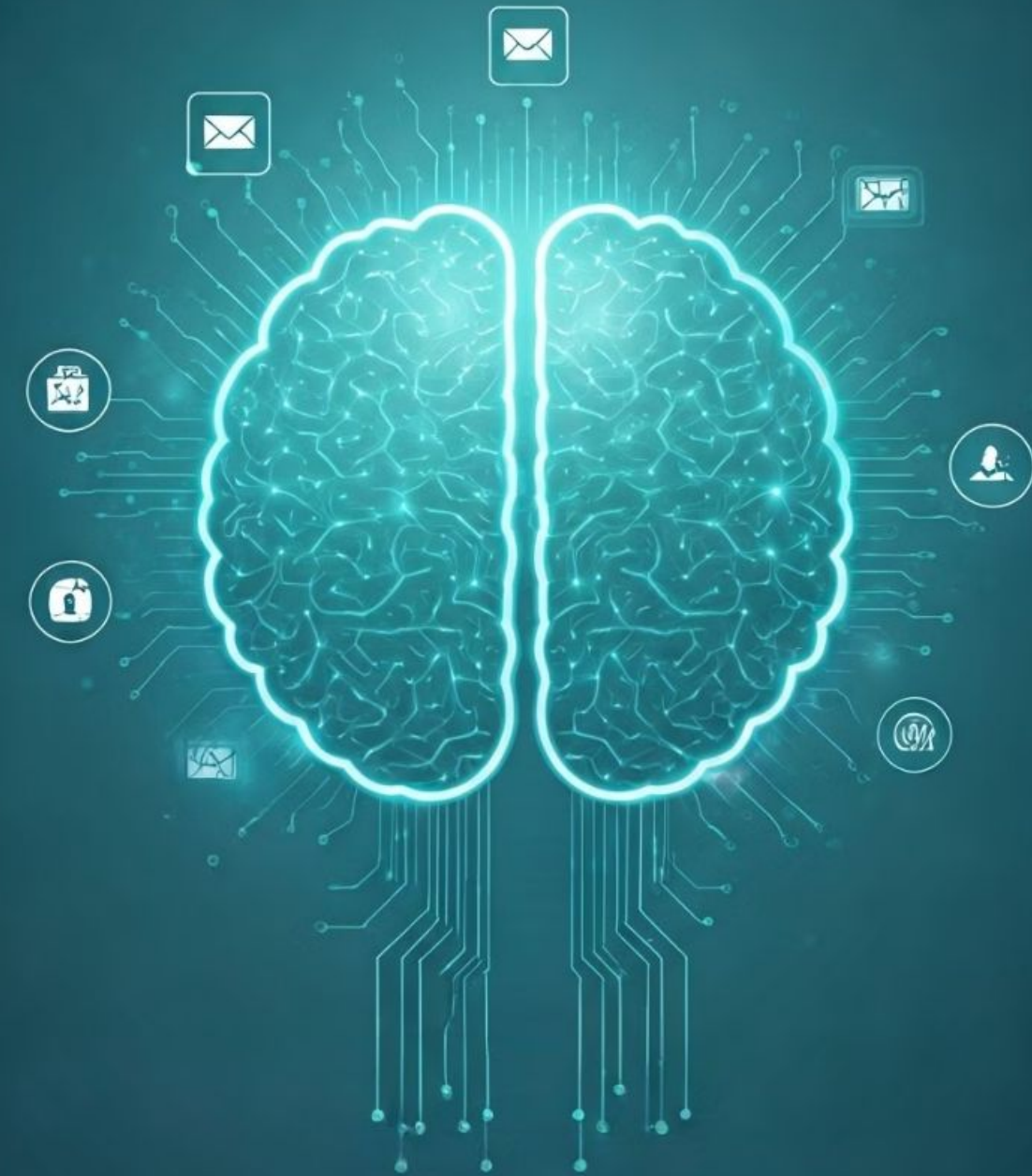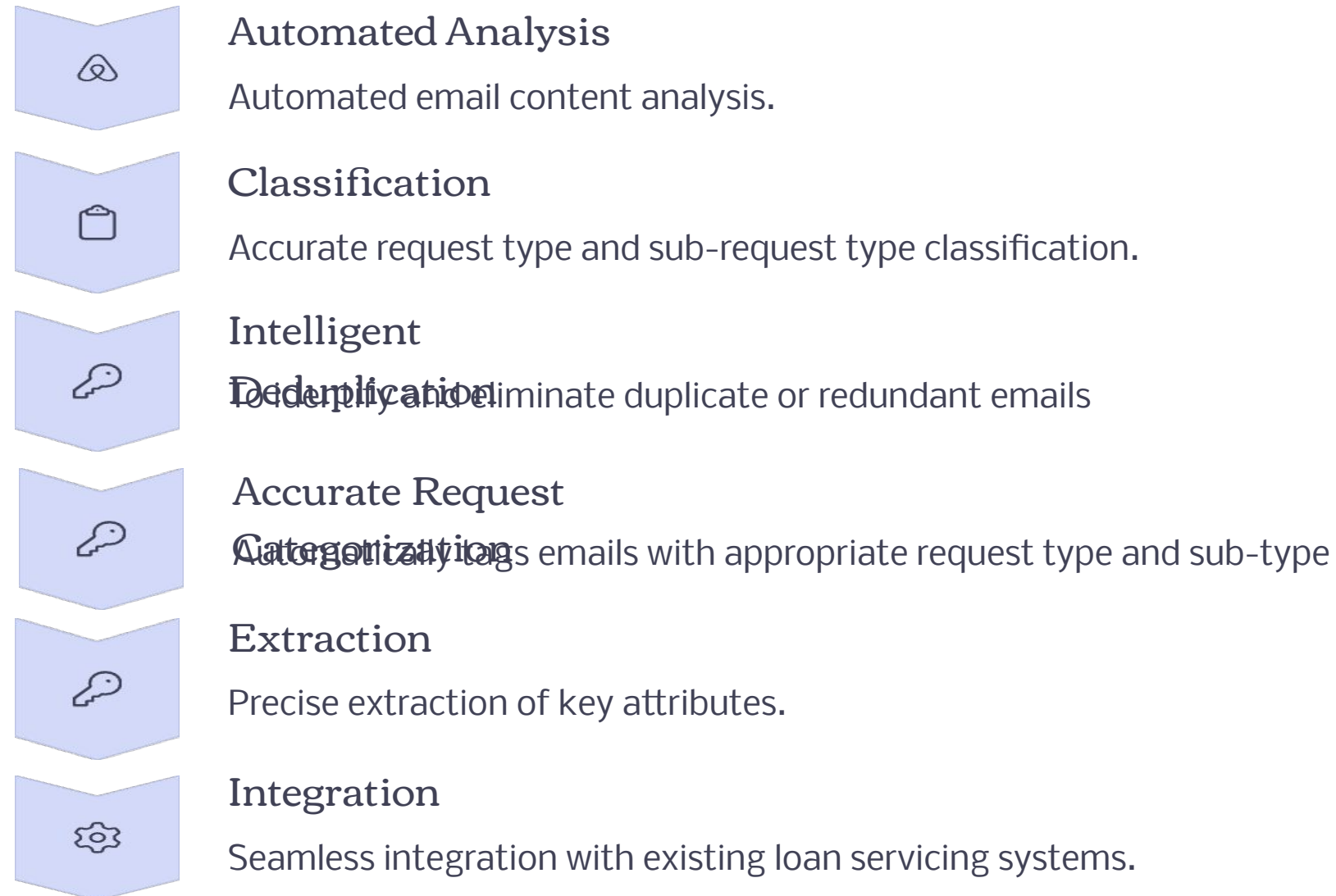Estimated annual cost in labor and lost productivity.

Backlogs, errors, and compliance risks plague the current system.

Proposed Solution: GEN AI Email Classification

To tackle the challenges of manual email triage process, we propose utilizing advanced Generative AI models for automated email classification. This cutting-edge solution will streamline the process, minimize errors, and enhance overall efficiency.

Automated Analysis

Automated email content analysis.

Classification

Accurate request type and sub-request type classification.

Intelligent Deduplication

Identify and eliminate duplicate or redundant emails

Accurate Request Categorization

Automatically tags emails with appropriate request type and sub-type

Extraction

Precise extraction of key attributes.

Integration

Seamless integration with existing loan servicing systems.

GEN AI Model Architecture

The GEN AI model architecture features a fine-tuned Large Language Model (LLM.

The system also integrates Optical Character Recognition (OCR) to extract and process textual data from scanned documents and image-based attachments, enabling end-to-end understanding of both email bodies and associated files.

LLM

Fine-tuned Large Language

Model (LLM): GENAI

OCR: Pytesseract

Training Data

Training data: historical loan

servicing emails with labeled

request types and attributes

Classification

Multi-stage classification.
Use multi-stage classification for request type and sub-request

type. Extract key attributes like Loan Number and Borrower Name.

Key Benefits: Enhanced Efficiency & Accuracy

Efficiency
Significantly reduce manual effort by automating the majority of email triage,

enabling faster, more accurate processing with minimal human intervention.

Improved Accuracy
Minimize errors and ensure consistent classification and customer satisfaction.

Reduced Costs
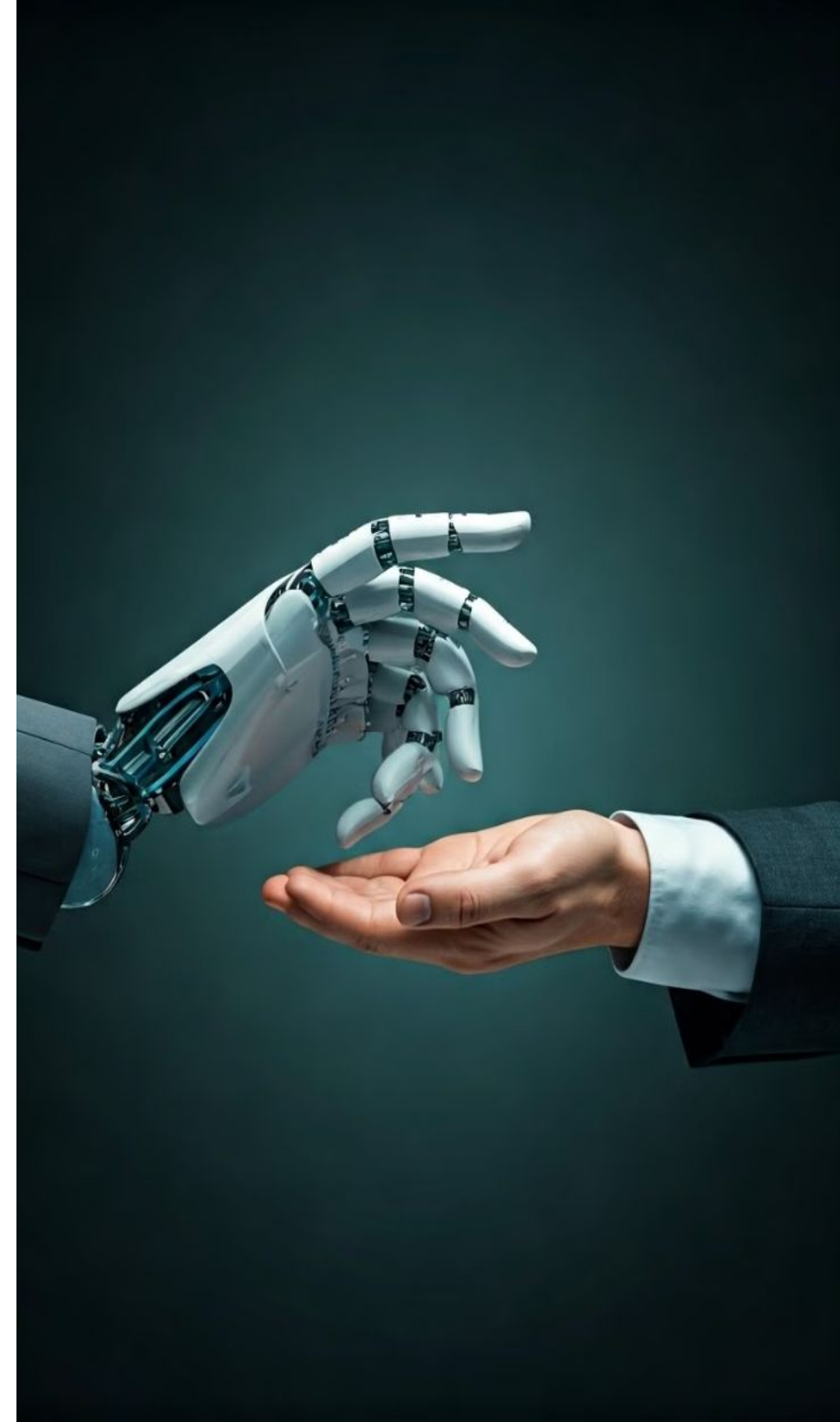Lower operational costs and increase profitability.

Scalability
Handles high volumes of emails effortlessly

Faster Turnaround Time
Accelerates request routing and resolution, improving overall response times.

Technology Stack & Integration

The GEN AI model architecture is built around a fine-tuned Large Language Model (LLM)

To ensure complete data coverage, the system integrates OCR capabilities using PyTesseract, enabling it to accurately extract text from scanned documents and image-based attachments such as PDFs and JPEGs.

This combination of LLM-driven understanding and OCR-based text extraction ensures the model can intelligently process both email content and embedded files–delivering end-to-end automation with high precision and reliability.

Python   OCR

LLM

Solution Overview

🔍 **Solution Overview: GEN AI–Powered Email Classification System**
Our solution automates the classification of loan servicing emails using a multi-stage pipeline that combines **Generative AI, OCR (PyTesseract),** and intelligent rule-based checks. It is designed to enhance processing efficiency, reduce manual workload, and ensure high accuracy.

🧠 1. **Email Parsing and Attachment Handling**
- Emails in .eml format are parsed to extract metadata, subject, body, and attachments.
- Attachments are saved locally for further processing.
- Both textual and image-based content are extracted for full context analysis.

📄 2. **OCR for Scanned Documents and PDFs**
- PyTesseract is used to extract text from image-based attachments (e.g., PNG, JPEG).
- PyMuPDF is used to extract text from PDF files, ensuring embedded and scanned data is fully captured.

🧹 3. **Preprocessing and Cleaning**
- Email content is cleaned to remove noise (e.g., formatting characters).
- Intelligent filtering detects and skips non-informative responses such as "Thank you" or "Noted".

♻️ **4. Duplicate Detection**
Emails are compared using a combination of:
Subject and body similarity (via difflib)
Attachment name matching
Time window proximity
Thank-you-only replies and duplicates are identified and filtered out using a smart scoring algorithm.
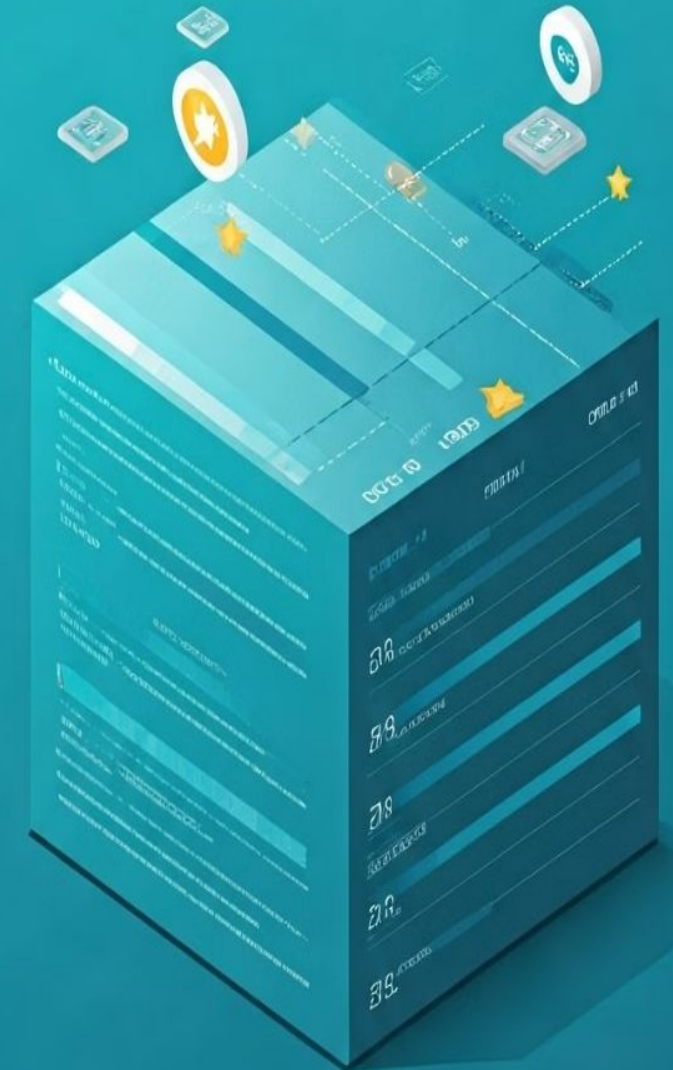
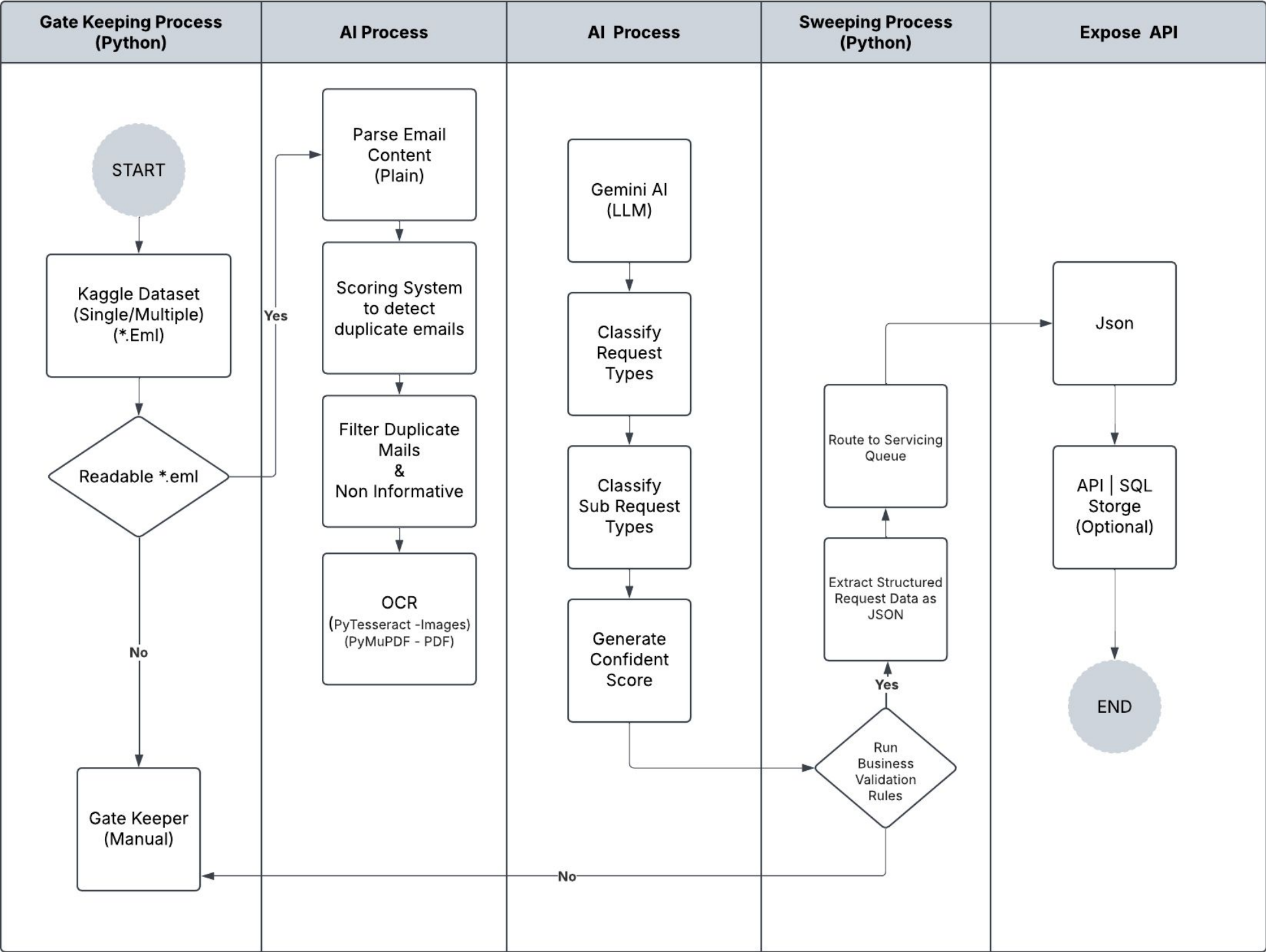# Solution Overview

🤖 **5. AI-Powered Classification with Gemini**

•The cleaned and enriched email context is passed to a fine-tuned Large Language Model (LLM) via the Gemini API.

•The model classifies the email into one of several predefined request types (e.g., Adjustment, Commitment Change, Money Movement).

•If applicable, a sub-request type is also identified.

•Confidence scores for both classifications are generated using structured JSON output.

📦 **6. Structured Request Extraction**

•Based on the predicted request type and sub-type, a final LLM prompt extracts structured details (e.g., Loan ID, Amount, Dates) in JSON format, enabling seamless downstream integration.

Solution Overview

Gate Keeping Process (Python)	AI Process	AI Process	Sweeping Process (Python)	Expose API

Gate Keeping Process (Python):
- START
- Kaggle Dataset (Single/Multiple) (*.Eml)
- Readable *.eml
  - Yes →
  - No → Gate Keeper (Manual)

AI Process:
- Parse Email Content (Plain)
- Scoring System to detect duplicate emails
- Filter Duplicate Mails & Non Informative
- OCR (PyTesseract -Images) (PyMuPDF - PDF)

AI Process:
- Gemini AI (LLM)
- Classify Request Types
- Classify Sub Request Types
- Generate Confident Score

Sweeping Process (Python):
- Route to Servicing Queue
- Extract Structured Request Data as JSON
- Run Business Validation Rules
  - Yes →
  - No → Gate Keeper (Manual)

Expose API:
- Json
- API | SQL Storge (Optional)
- END

Sample Output

```
/kaggle/input/textimage/Email_15_MOR6719.eml

Saved attachment: email/attachments/Attachment_15_MOR6719.jpeg
▼ root
    Adjustment 0.1
    AU Transfer 0.7
    Closing Notice 0.1
    Commitment Change 0.1
    Fees Payment 0.1
    Money Movement-Inbound 0.95
    Money Movement-Outbound 0.1
▼ root
    Principal 0.95
    Interest 0.01
    Principal+Interest 0.02
    Principal+Interest+Fee 0.02
```

Output shows classification confidence score

Sample Output

```
request_type "Money Movement-Inbound"
subrequest_type "Principal"
▼ details
    date "24-Mar-2025"
    borrower "JP MORGAN CHASE"
    facility_type "REVOLVER"
    reference "JP7038"
    deal_cusip "97007EB39"
    deal_isin "US97007EB39"
    lender_mei "US1L044730"
    effective_date "22-Apr-2025"
    total_amount null
    currency "USD"
    previous_global_principal_balance "55,549,676.00"
    new_global_principal_balance "5,549,676.00"
    loan_origination_date "22-Jan-2025"
    repricing_date "22-May-2025"
    lender_share "10,000,000.00"
    previous_lender_share_principal_balance "11,009,707.00"
    new_lender_share_principal_balance "1,009,707.00"
    submission_amount "10,000,000.00"
    for "CITI GROUP"
    to "CITI GROUP"
    aba_number "021000021"
    account_number "XXXXXXX866341"
```

Output shows the extracted details from the email .

NOTE: the X's in the account number etc are from the email. not from the model

Future Scope
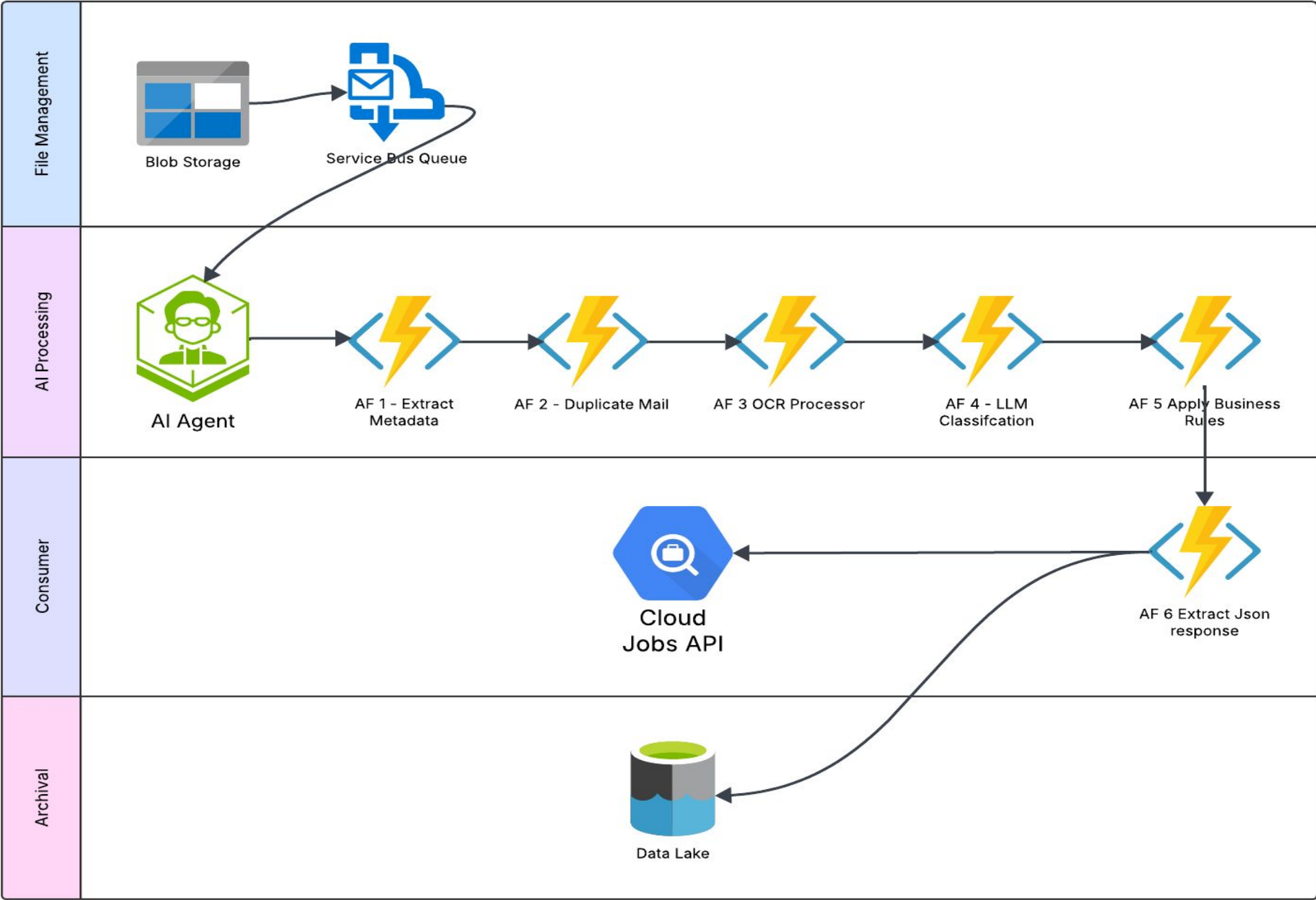
✅ Key Benefits

•Automates email triage with minimal human intervention

•Handles both structured and unstructured data (including images/PDFs)

•Removes redundancy through duplicate and thank-you detection

•Provides accurate, explainable classifications with confidence scores

•Scalable, adaptable, and ready for integration with loan servicing workflows

•Reduces manual errors and inconsistencies by standardizing request interpretation

•Future-proof and extensible architecture that allows easy inclusion of new request type
 workflows

•Enables data-driven insights by converting unstructured emails into analyzable structu

Performance Metrics & Results

The success of the GEN AI email classification system will be measured using several key performance metrics. These metrics provide a clear indication of the system's effectiveness in improving accuracy, efficiency, and cost savings.

Accuracy
Achieve >95% classification accuracy on test dataset

Efficiency
Reduce email processing time by 75%

Cost
Generate in annual cost savings

Real-World Example and Use Cases

Specific Use Cases

- Payment Deferral Requests

- Account Update Requests

- Loan Payoff Requests

- Interest Rate Change Inquiries

- Loan Renewal or Rollover Instructions

- Dispute or Error Resolution Requests

Conclusion: Transforming Loan Servicing with GEN AI

By leveraging the power of Generative AI and OCR, our solution transforms a

traditionally manual, error-prone email triage process into a streamlined,

intelligent system. It not only accelerates response times but also enhances

GEN AI-powered email classification revolutionizes loan servicing by accuracy, scalability, and adaptability — making it an essential tool for modernizing

It significantly improves outcomes.

loan servicing operations.

It strengthens competitive advantage and paves the way for AI innovations.

Acc
ura

2

Effi
cien
cy

1

3

Customer
Satisfacti
on
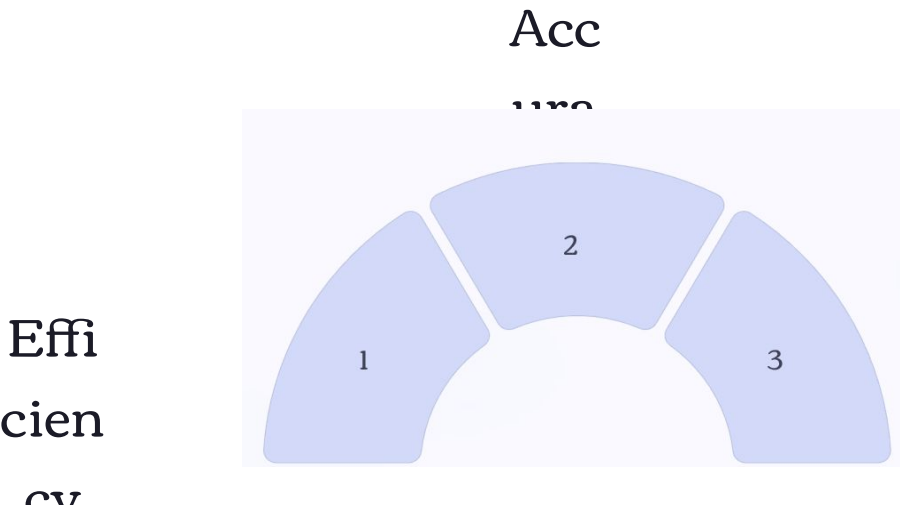
Team

Chirag Jasuja

Rajesh Karnoon

Subham Bhattacharjee

Kumaran Arulprakasam

Mahesh Patil