POC Request: AI-Powered Summarization of Unstructured Data

1. Overview

The objective of this Proof of Concept (POC) is to evaluate and implement an AI-driven solution that can process large volumes of unstructured data and generate concise, accurate, and business-relevant summaries. Currently, our teams receive information from multiple heterogeneous sources in varying formats — including free-form text, tables without headers, embedded images, copied snapshots, and other non-standard structures. Manually analyzing such data is time-consuming, inconsistent, and prone to human error. This POC aims to automate the extraction of meaningful insights and provide summarized information that can support faster and more accurate decision-making.

2. Business Problem

Our organization regularly deals with unstructured data that requires contextual understanding to derive business insights. Key challenges include: • Unstructured Inputs: Data exists in diverse formats, such as images, tables without headers, free-text notes, and mixed content. • High Manual Effort: Current processes require significant human intervention to parse, understand, and summarize content. • Inconsistent Interpretations: Different teams interpret unstructured data differently, leading to potential discrepancies. • Time Sensitivity: In critical scenarios, delays in analyzing large data sets impact operational efficiency and decision timelines.

3. Objective of the POC

The goal is to assess the feasibility of using AI-based summarization models to: • Automatically parse, process, and interpret unstructured data across multiple formats. • Identify key entities, metrics, and contextual relationships without requiring predefined data schemas. • Generate concise, accurate summaries tailored to specific business objectives. • Reduce manual dependencies and improve response times in time-sensitive workflows.

4. Scope

The POC will focus on: • Parsing multi-format data sources (free-form text, tabular data including incomplete or headerless tables, and embedded images). • Applying AI-driven techniques for text extraction, normalization, entity identification, and summarization. • Delivering summarized outputs aligned with stakeholder requirements.

5. Approach

1. Data Ingestion & Preprocessing: - Collect representative unstructured data samples. - Apply preprocessing techniques such as OCR for images, parsing for tables, and cleaning for raw inputs. 2. Model Evaluation: - Evaluate traditional AI/ML models for NER, table parsing, and summarization. - Compare approaches for accuracy, scalability, and performance. 3. Summarization Framework: - Combine rule-based techniques with AI-driven models to extract key

entities and relationships, preserve context, and generate summaries. 4. Validation & Benchmarking: - Define evaluation metrics such as precision, recall, summary accuracy, and time savings. - Gather business stakeholder feedback to validate outcomes.

6. Expected Benefits

• Operational Efficiency: Significant reduction in manual analysis efforts. • Consistency & Accuracy: Standardized interpretation across teams. • Faster Decision-Making: Summaries delivered in near real-time. • Scalability: Ability to handle growing data volumes without increasing resource requirements.

7. Deliverables

• Summarized Reports: AI-generated summaries of unstructured datasets. • POC Findings Document: Detailing performance metrics, limitations, and recommendations. • Implementation Roadmap: Next steps for enterprise-wide adoption if POC is successful.

8. Risk & Mitigation

Risk	Impact	Mitigation
Model accuracy variation	Medium	Use a hybrid approach (AI + rule-based)
Data complexity	High	Iterative refinement of preprocessing and extraction
Stakeholder alignment	Medium	Regular reviews and feedback loops

9. Request

Approval is requested to proceed with this POC to evaluate AI-based summarization capabilities for unstructured data. A successful POC will form the foundation for scaling this solution across the organization.