TymeBank Case Study Narrative

In this case study, the researcher was tasked with designing, implementing, and monitoring an end-to-end document extraction solution built around a centralized HTTP API to address inefficiencies in TymeBank's current data pipeline. At present, multiple business teams – including Loans, Risk, Operations, and Legal – have developed their own OCR workflows, resulting in fragmented systems, inconsistent accuracy, and redundant processing across teams.

To address this, the researcher proposed a unified pipeline hosted on the Databricks platform. The system leverages a **hybrid document parsing architecture**, combining **production-grade OCR engines** (e.g., AWS Textract, Tesseract) with lightweight **Python-based extraction libraries** (e.g., PyMuPDF, python-docx, pandas) depending on document type and quality. Extracted text is then passed through a **two-stage LLM pipeline**: the first LLM (Claude 3.5 Sonnet) performs structured field extraction, flags uncertainties, and outputs schema-aligned JSON; the second LLM (GPT-4 Turbo) verifies each field, resolves flagged issues, and finalizes a clean, hallucination-free JSON output for downstream consumption.

For monitoring and quality assurance, the researcher designed a performance evaluation framework based on four key metrics: **latency, error rate, field-level accuracy, and drift/degradation detection**. These are measured through a combination of **weekly audits** (sampling output by segment) and **monthly evaluations** (comparing LLM outputs against ground-truth annotations across document types and business use cases).

A one-week experimentation plan supports this rollout by benchmarking the pipeline's accuracy and robustness. This includes targeted testing across varied input formats and schemas, tracking how well the LLM generalizes across domains, and refining prompt design to improve JSON output quality. The results of this phase will guide further iteration and scaling across TymeBank's internal teams.