SOCAR Historical Documents AI

Intelligent OCR & RAG System for Oil & Gas Archives

Transforming 28 Historical Documents into Searchable Knowledge

Team BeatByte

Ulvi Bashirov | Samir Mehdiyev | Ismat Samadov

! The Problem

PDF

Inaccessible Archives

Decades of valuable historical documents locked in PDF format, impossible to search

ABC

Multi-Language Barrier

Documents in Azerbaijani, Russian, and English with complex Cyrillic text

TIME

Time-Consuming Research

Manual document review takes hours to find specific information

How can we unlock institutional knowledge trapped in historical documents?

* Our Solution

Vision-Language OCR

State-of-the-art Llama-4-Maverick model extracts text from scanned documents with **87.75% accuracy**, preserving Cyrillic characters perfectly

Semantic Search

BAAI/bge-large embeddings + Pinecone vector database enable instant retrieval across **1,128 document chunks**

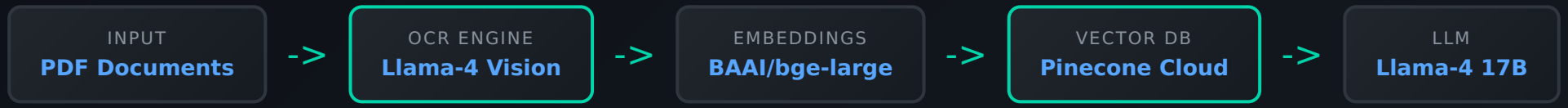
RAG-Powered Q&A

Natural language questions answered with relevant context and **source citations** for verification

Production-Ready API

FastAPI backend with Docker deployment, health monitoring, and interactive web interface

System Architecture



OCR Pipeline

PDF Upload -> PyMuPDF (100 DPI) -> Vision LLM

Image Detection -> Markdown Output

RAG Pipeline

User Question -> Embed Query -> Top-3 Retrieval

Context Building -> LLM + Citations

+ Technology Stack

L Llama-4-Maverick 17B Vision & Language Model	B BAAI/bge-large-en 1024-dim Embeddings	P Pinecone Cloud Vector Database
F FastAPI Async REST API	M PyMuPDF PDF Processing	D Docker Containerization

API Endpoints

POST `/ocr`
Extract text from uploaded PDF with image detection

POST `/llm`
RAG-based Q&A with source citations

GET `/health`
Service health check and vector count

% Benchmark Results

We rigorously tested **3 OCR models**, **7 RAG configurations**, and **3 LLMs** to optimize performance

OCR Model Comparison

Model	Character Success Rate	Word Success Rate	Speed (12 pages)	Type
GPT-4.1	88.12%	67.44%	199s	Closed
Llama-4-Maverick 17B [Selected]	**87.75%**	**61.91%**	**75s**	Open
Phi-4-multimodal	Failed			Open

Selected Llama-4: Only 0.37% accuracy loss vs GPT-4.1, but **2.7x faster** and **open-source**

@ RAG Optimization Results

Configuration	Answer Quality	Citation Rate	Response Time
Citation-focused + Vanilla k3 [Selected]	**55.67%**	**73.33%**	**3.61s**
Few-shot + Vanilla k3	45.70%	40.00%	2.17s
Baseline + Vanilla k3	39.65%	20.00%	2.28s
MMR Retrieval	34.60%	6.67%	2.53s

Key Insight: Simple Beats Complex

Vanilla retrieval outperforms MMR reranking by **+21%**. Top-3 beats Top-5 by **+20%**

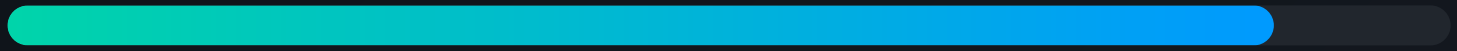
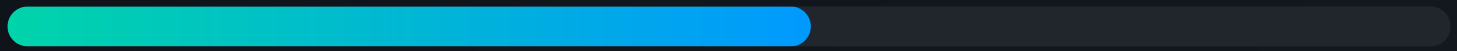
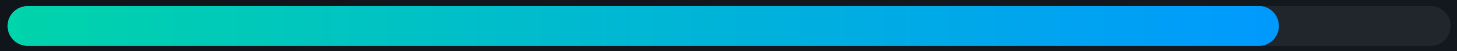
Citation-Focused Prompting

Custom Azerbaijani prompt improves quality by **+16%** and citation rate by **+53%**

^ Performance Metrics

87.75%	55.67%	73.33%	3.6s
OCR ACCURACY	ANSWER QUALITY	CITATION RATE	RESPONSE TIME

Estimated Hackathon Score

OCR Quality (50%)		43.9 / 50
LLM Quality (30%)		16.7 / 30
Architecture (20%)		20 / 20
TOTAL SCORE		**440.6 / 500**

& Key Technical Decisions

What We Did

> **Open-source Llama** over proprietary GPT-4
> **Top-3 retrieval** - more context confused the LLM
> **Vanilla retrieval** - simple beats complex reranking
> **Citation-focused prompt** in Azerbaijani
> **BAAI embeddings** - 25% better than multilingual
> **600-char chunks** with 100-char overlap

What We Avoided

> **MMR/Reranking** - 21% worse performance
> **Top-5+ retrieval** - information overload
> **Few-shot prompting** - inconsistent results
> **Multilingual embeddings** - underperformed
> **Complex architectures** - kept it simple
> **Closed-source models** - for transparency

"Every decision was validated through rigorous benchmarking across 3 Jupyter notebooks"

> Live Demo Features

[^] PDF Upload & OCR
Drag & drop any PDF to extract text with image detection. Results in markdown format.

[?] Interactive Q&A Chat
Ask questions in Azerbaijani, Russian, or English. Get answers with source citations.

[i] Source Citations
Every answer includes document name, page number, and relevant excerpt for verification.

[=] Swagger Documentation
Full API documentation at /docs with interactive testing capabilities.

Web UI: **localhost:8000** | API Docs: **/docs**

= Deliverables

28	1,128	3	100%
PDFS PROCESSED	VECTOR CHUNKS	BENCHMARK NOTEBOOKS	OPEN SOURCE

Code & Infrastructure

> FastAPI application (505 lines)
> Data ingestion pipeline
> Parallel processing (4x speedup)
> Docker + Docker Compose
> Health monitoring
> Interactive web UI

Documentation & Analysis

> 8 comprehensive markdown docs
> VLM OCR benchmark notebook
> RAG optimization notebook
> LLM comparison notebook
> Sample questions & answers
> Deployment guide

Thank You!

SOCAR Historical Documents AI System

Transforming archives into accessible, searchable knowledge

Team BeatByte

Ulvi Bashirov | Samir Mehdiyev | Ismat Samadov

87.75%	440.6	100%	3.6s
OCR ACCURACY	EST. SCORE / 500	OPEN SOURCE	RESPONSE TIME

Questions? Let's Demo!