Uzbek Linguistic Pipeline

AI-Powered Digitization for a Low-Resource Language

- Enterprise-grade platform transforming historical printed dictionaries into structured NLP databases.
- Powered by Google Gemini 3.1 Pro via Vertex AI.
- Laying the foundation for next-generation Uzbek TTS, G2P, and AI models.

The Low-Resource Challenge & The Formatting Barrier

The Macro Challenge

35M+ global speakers, yet severely low-resource in AI corpora.

The Micro Challenge

O'T I [öт] *n.* (*bot.*) grass, herb; hay; pasture.
Etym.: < Ar. عُشْب ('ushb); Pers. گیاه (giyāh); Russ. трава́ (travá).

O'T II [öт] *n.* (*anat.*) gall bladder; bile.
Etym.: < Ar. عُشْب ('ushb); **II**, Pers. گیاه (giyāh); Russ. трава́ (travá).

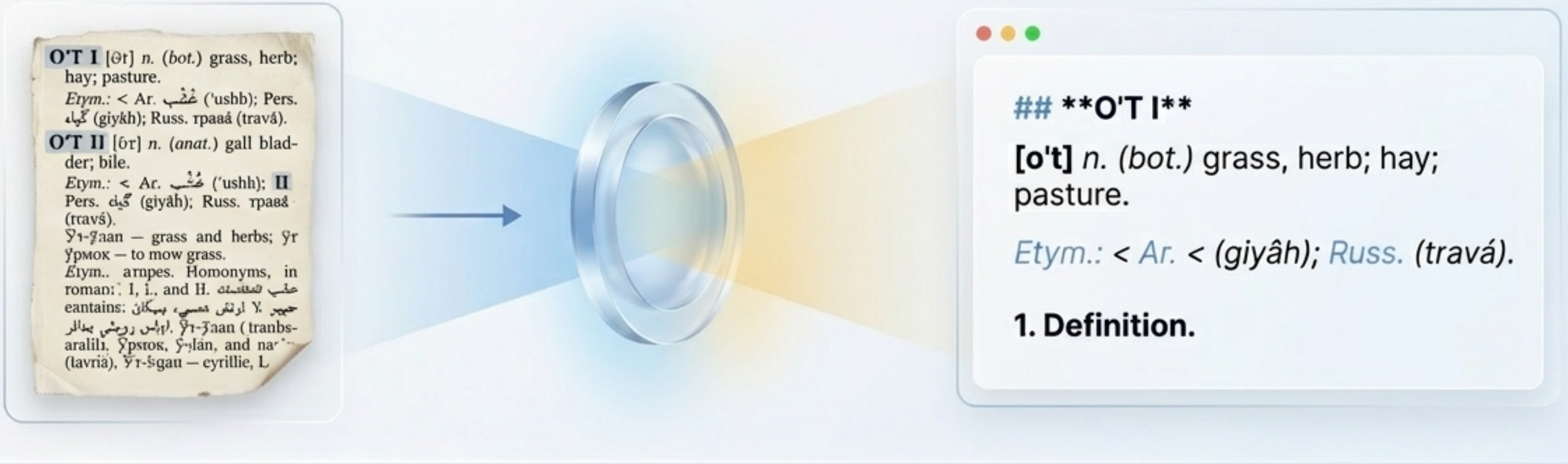
Ўт-ўлан — grass and herbs; ўт ўрмоқ — to mow grass.

Etym.. ятпрes. Homonyms, in romani. I, i., and II. عشب تحتوصلت cantains: حيبر ، لا ارنش دمسي، بسيكان اولس رويشي بنافر. Ўт-ўлан (tranbs-aralii), Ўрмоқ, ў-jlān, and nariv, (tavriā), Ўт-ўган — cyrillic, Latin,

Traditional OCR destroys structural hierarchy. It fails to parse:

- Bold text (Headwords)
- Italic text (Etymologies: Arabic, Persian, Russian)
- Roman numerals (Homonyms like O'T I, O'T II)
- Mixed character sets (Cyrillic, Latin, Arabic)

Shifting from Text Extraction to Semantic Extraction



Core Concept

Using Multimodal Large Language Models as linguistic OCR agents.

Mechanism

The AI interprets the visual layout and outputs strictly formatted Markdown. This preserves the semantic structure (**Headword**, *etymology*, *1. Definition*) required for relational database mapping.

Enterprise Architecture and Tech Stack

Infrastructure & AI

- Google Cloud Platform (GCP)
- Compute Engine & Vertex AI
- Gemini 3.1 Pro Preview

Backend System

- Python 3.10+
- Django 5.2
- PostgreSQL

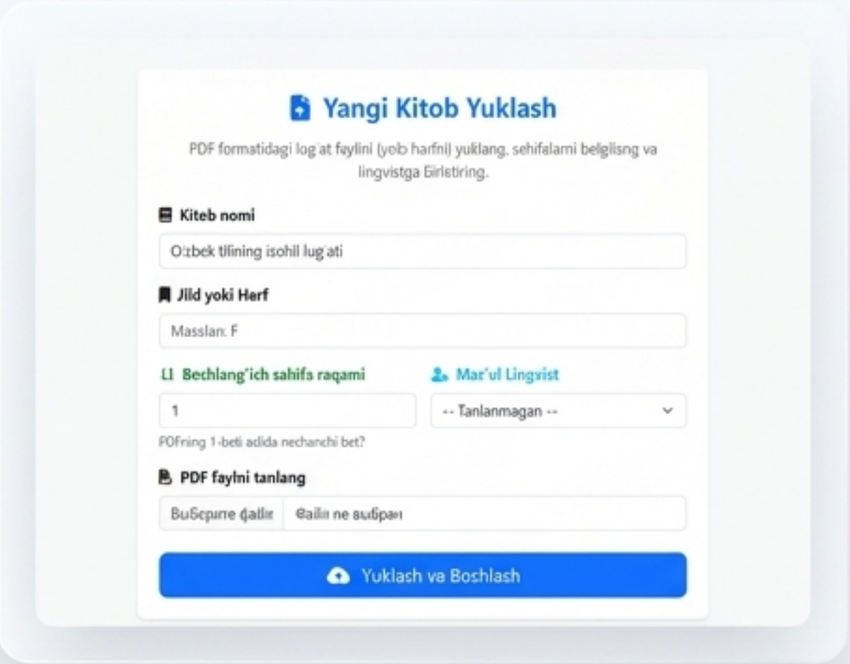
Asynchronous Engine

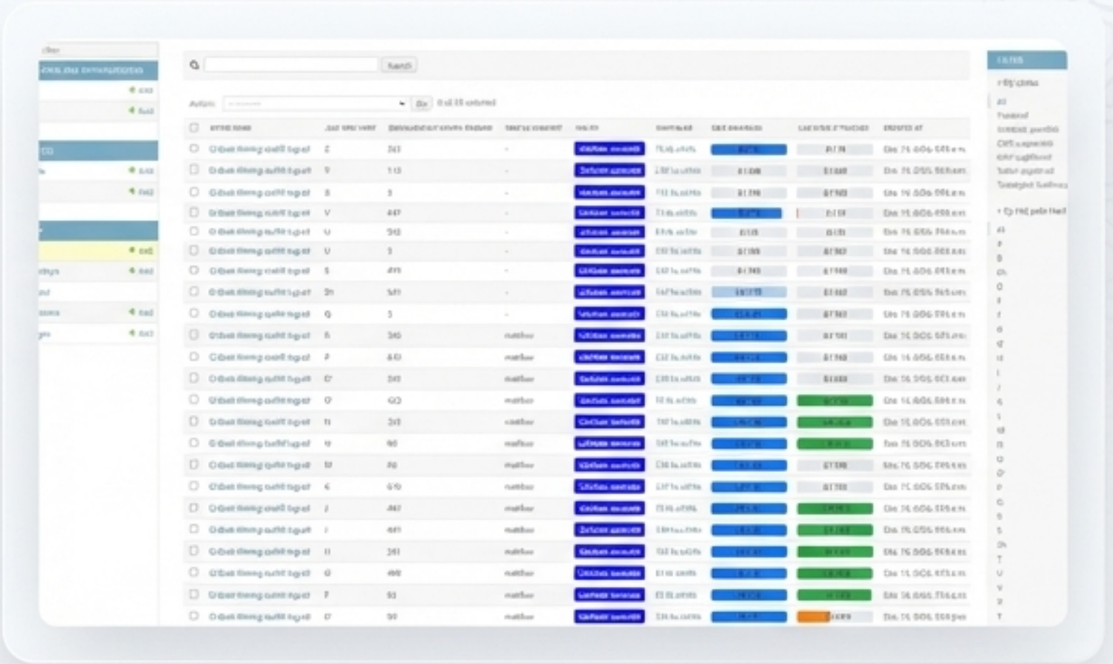
- Celery & Redis
- Processes thousands of jobs without freezing the web server.

Frontend UI

- HTML5 & Bootstrap 5
- Chart.js
- Glassmorphism UI components

Pipeline Steps 1 & 2: Ingestion and Asynchronous AI OCR





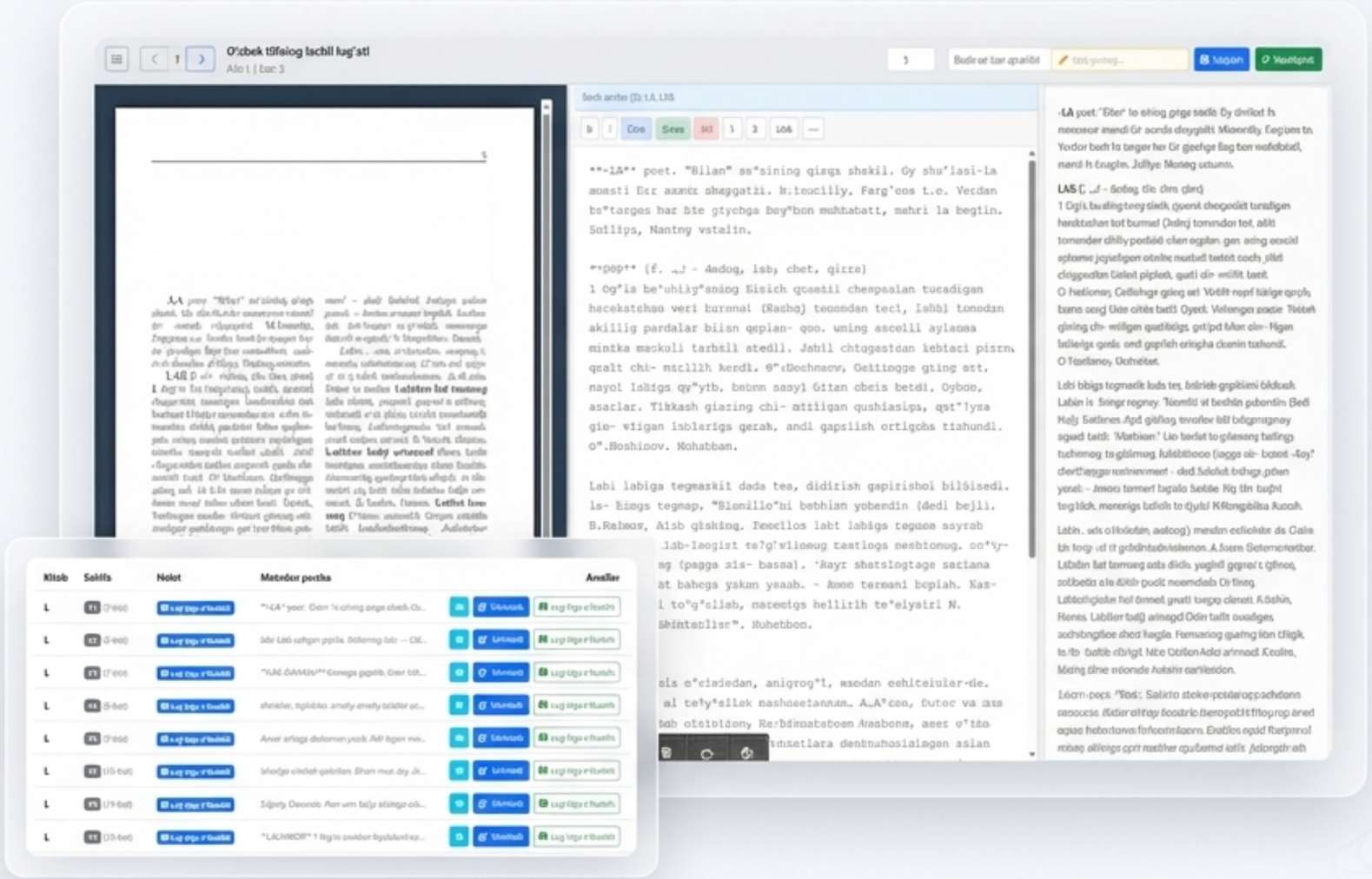
Step 1: Document Upload

Admins upload 500+ page PDFs. Background Celery workers extract high-resolution images.

Step 2: Multimodal OCR

Parallel processing via Vertex AI. Gemini 3.1 Pro analyzes images with strict system prompts to generate raw Markdown.

Pipeline Step 3: Human-in-the-Loop Verification



Workspace Features

Secure portal for professional linguists to prevent overlapping work.

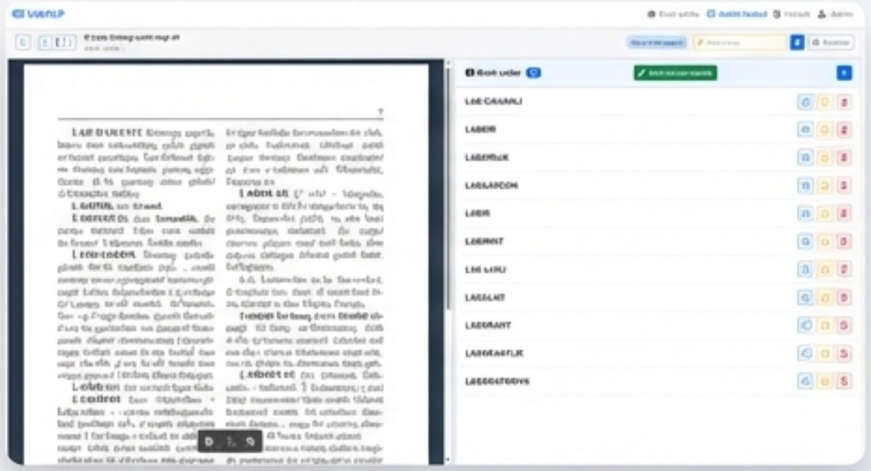
Split-Screen UI

Side-by-side view of the original scanned image and AI-generated Markdown. Enables rapid fixing of minor hallucinations and 100% accuracy verification.

Pipeline Steps 4, 5 & 6: NLP Parsing to Public API

Verified Markdown

Python Regex



Public API

Step 4 & 5: NLP Parsing

Python Regular Expressions parse bolded patterns (**WORD***) and handle Uzbek-specific characters (‘ tutuq belgisi’). Dynamically processes homonyms via Roman numerals into DictionaryEntry objects.

Step 6: Public APIs

Structured data is indexed for sub-second search latency and public display.

Tailored Ecosystem Access Control

Public Users

- Access to live metrics displays
- Chart.js interactive analytics
- Sub-second search

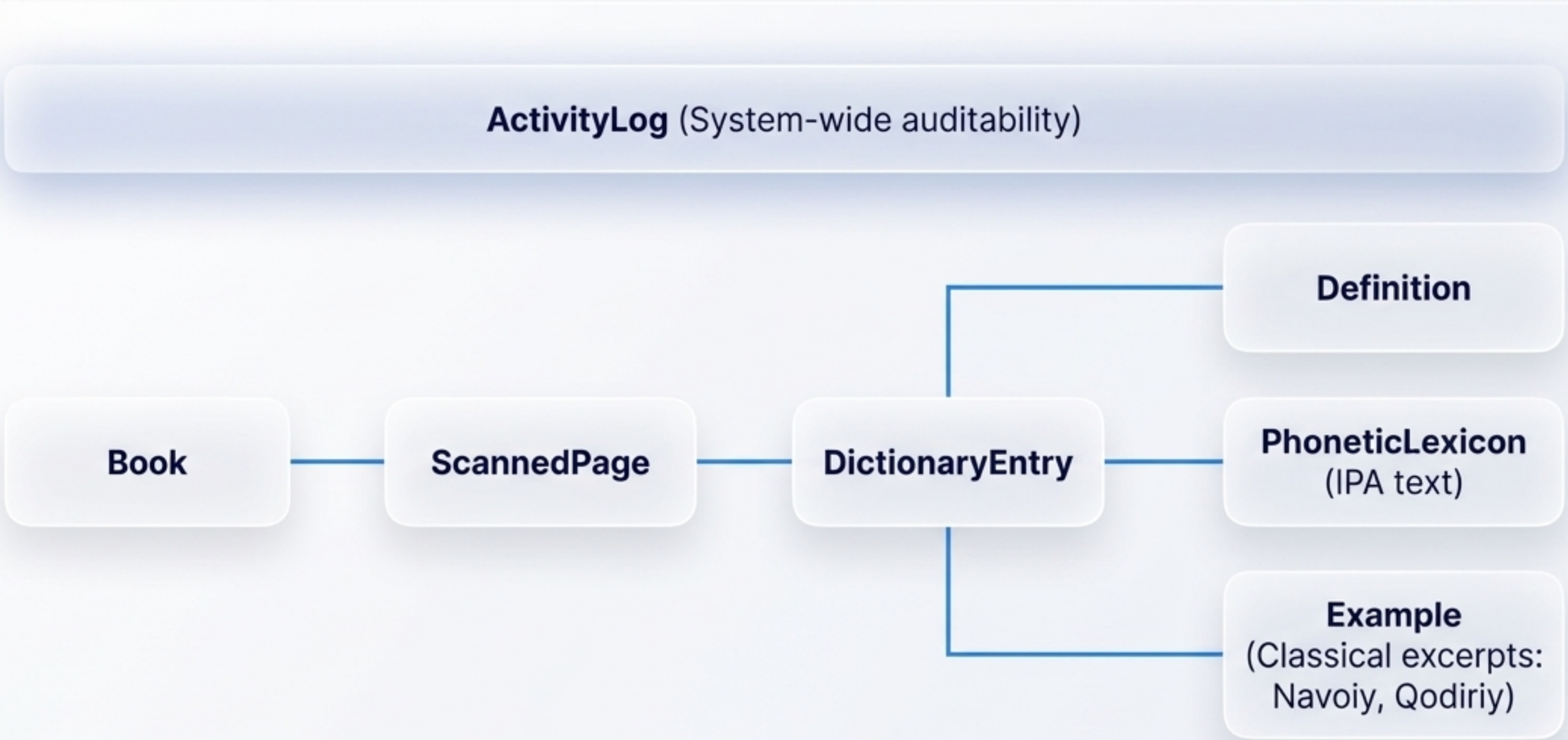
Linguists

- Secure Markdown editor
- Contextual verification
- Internal editor notes for printing errors

SuperAdmins

- Complete CRUD access
- Batch operations for Celery OCR tasks
- Real-time job monitoring via django_celery_results

Deep Relational Linguistic Schema



Live Production Metrics (As of March 2026)

Scale Metrics

28
Dictionary Volumes Migrated

5,190
Total Scanned Pages

Processing Pipeline

Processing Progress

3,881
Pages successfully processed by AI

1,312
Pages human-verified

Data Yield

11,852
Unique Headwords Extracted

Dataset Granularity and Activity Logging

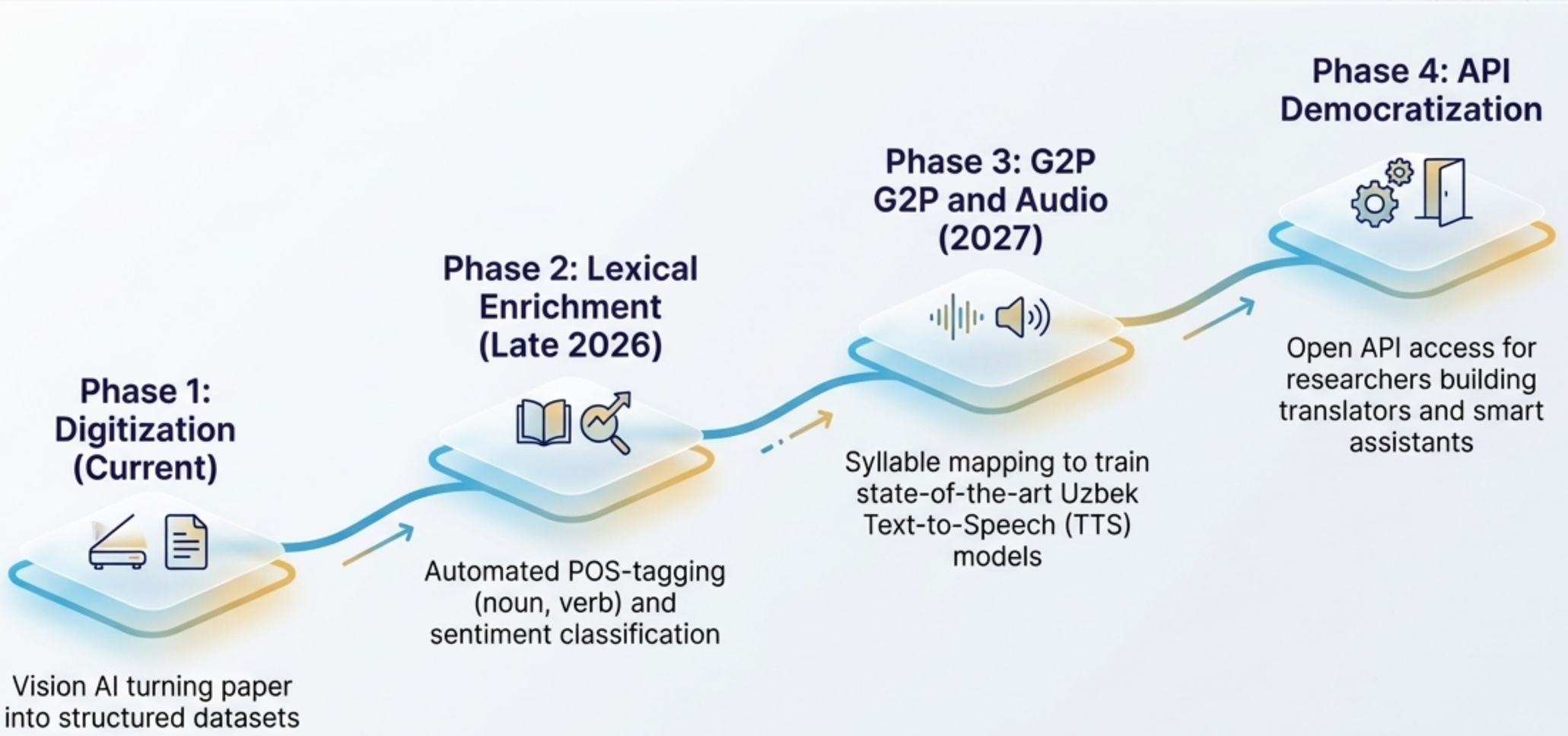
System Activity

Over **11,632** automated operations tracked by the internal audit logger, ensuring complete data provenance.

Synchronous Volume Progression

100% OCR & Parsed	Near 100% Parsed	100% OCR / Pending Parsing	Currently in OCR Queue
Volumes A, E, G, G‘, H, I, L, N	Volumes O, O‘, F, J	Volumes B, K, M, P, Q, R, Sh, Z	Volumes S (345 pages) Volumes T (547 pages) Volumes T (547 pages) Volumes Y (208 pages) Volumes X (110 pages) Volumes U (95 pages)

The NLP Expansion Roadmap (2026–2027+)



Ensuring a Low-Resource Language Thrives in the AI Era

The Uzbek Linguistic Pipeline bridges centuries of lexicography with Google Vertex AI—empowering human experts to build the foundational data infrastructure for the next generation of intelligent systems.