

# 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** [öt] *n.* (*bot.*) grass, herb; hay; pasture.

*Etym.:* < Ar. عُشْب ('ushb); Pers. گیاه (giyāh); Russ. травá (travá).

**O'T II** [öt] *n.* (*anat.*) gall bladder; bile.

*Etym.:* < Ar. عُشْب ('ushb); **II**, Pers. گیاه (giyāh); Russ. травá (travá).

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

*Etym.:* ятipes. Homonyms, in romanı̄. I, i., and II. عشب تحاضمت حبير, لا ارنش دمسي, بيبكان cantains: اولس رويني بنافر. Ўт-ўлан (transaralii), ўрмок, ў-їлан, and nariv, (tavriā), Ўт-їган — 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

**O'T I** [θɪ] *n.* (bot.) grass, herb; hay; pasture.

*Etym.:* < Ar. عُشْب ('ushb); Pers. گیاه (giyâh); Russ. травá (travá).

**O'T II** [θɪ] *n.* (anat.) gall bladder; bile.

*Etym.:* < Ar. عُشْب ('ushh); II Pers. گیاه (giyâh); Russ. травá (travá).

Ṣī-ṣāan — grass and herbs; ṣr ṣpmox — to mow grass.

*Etym.* araps. Homonyms, in roman: I, i., and H. عشب العنقبيات

عنتاين: ي. اوتش قشبي، بيمكان

جهر. Ṣī-ṣāan (trans-aralili, ṣpstos, ṣ-ṣlān, and nar

(lavriá), Ṣr-ṣgau — cyrillie, L



## \*\*O'T I\*\*

**[o't]** *n.* (bot.) grass, herb; hay; pasture.

*Etym.:* < Ar. < (giyâh); Russ. (travá).

**1. Definition.**

## 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

**Yangi Kitob Yuklash**

PDF formatidagi log at faylini (yob harfni) yuklang, sahifalarni belgilsng va lingvistga Birlsiring.

**Kitob nomi**  
O'zbek tiling isohil lug'ati

**Jild yoki Herf**  
Masslar. F

**LI Bechlang'ich sahifa raqami** **Mar'ul Lingvist**  
1 -- Tanlanmagan --

PDFning 1-beti adida nechanchi bet?

**PDF faylni tanlang**  
BuSopume qalir Galir ne sulpan

**Yuklash va Boshlash**



ID	Kitob nomi	Jild yoki Herf	Bechlang'ich sahifa raqami	Mar'ul Lingvist	Status	Yaratilgan	Yilgan	Yilgan	Yilgan
1	O'zbek tiling isohil lug'ati	F	1	Mar'ul Lingvist	Yaratilgan	2024-01-01	2024-01-01	2024-01-01	2024-01-01
2	O'zbek tiling isohil lug'ati	F	2	Mar'ul Lingvist	Yaratilgan	2024-01-01	2024-01-01	2024-01-01	2024-01-01
3	O'zbek tiling isohil lug'ati	F	3	Mar'ul Lingvist	Yaratilgan	2024-01-01	2024-01-01	2024-01-01	2024-01-01
4	O'zbek tiling isohil lug'ati	F	4	Mar'ul Lingvist	Yaratilgan	2024-01-01	2024-01-01	2024-01-01	2024-01-01
5	O'zbek tiling isohil lug'ati	F	5	Mar'ul Lingvist	Yaratilgan	2024-01-01	2024-01-01	2024-01-01	2024-01-01
6	O'zbek tiling isohil lug'ati	F	6	Mar'ul Lingvist	Yaratilgan	2024-01-01	2024-01-01	2024-01-01	2024-01-01
7	O'zbek tiling isohil lug'ati	F	7	Mar'ul Lingvist	Yaratilgan	2024-01-01	2024-01-01	2024-01-01	2024-01-01
8	O'zbek tiling isohil lug'ati	F	8	Mar'ul Lingvist	Yaratilgan	2024-01-01	2024-01-01	2024-01-01	2024-01-01
9	O'zbek tiling isohil lug'ati	F	9	Mar'ul Lingvist	Yaratilgan	2024-01-01	2024-01-01	2024-01-01	2024-01-01
10	O'zbek tiling isohil lug'ati	F	10	Mar'ul Lingvist	Yaratilgan	2024-01-01	2024-01-01	2024-01-01	2024-01-01
11	O'zbek tiling isohil lug'ati	F	11	Mar'ul Lingvist	Yaratilgan	2024-01-01	2024-01-01	2024-01-01	2024-01-01
12	O'zbek tiling isohil lug'ati	F	12	Mar'ul Lingvist	Yaratilgan	2024-01-01	2024-01-01	2024-01-01	2024-01-01
13	O'zbek tiling isohil lug'ati	F	13	Mar'ul Lingvist	Yaratilgan	2024-01-01	2024-01-01	2024-01-01	2024-01-01
14	O'zbek tiling isohil lug'ati	F	14	Mar'ul Lingvist	Yaratilgan	2024-01-01	2024-01-01	2024-01-01	2024-01-01
15	O'zbek tiling isohil lug'ati	F	15	Mar'ul Lingvist	Yaratilgan	2024-01-01	2024-01-01	2024-01-01	2024-01-01
16	O'zbek tiling isohil lug'ati	F	16	Mar'ul Lingvist	Yaratilgan	2024-01-01	2024-01-01	2024-01-01	2024-01-01
17	O'zbek tiling isohil lug'ati	F	17	Mar'ul Lingvist	Yaratilgan	2024-01-01	2024-01-01	2024-01-01	2024-01-01
18	O'zbek tiling isohil lug'ati	F	18	Mar'ul Lingvist	Yaratilgan	2024-01-01	2024-01-01	2024-01-01	2024-01-01
19	O'zbek tiling isohil lug'ati	F	19	Mar'ul Lingvist	Yaratilgan	2024-01-01	2024-01-01	2024-01-01	2024-01-01
20	O'zbek tiling isohil lug'ati	F	20	Mar'ul Lingvist	Yaratilgan	2024-01-01	2024-01-01	2024-01-01	2024-01-01

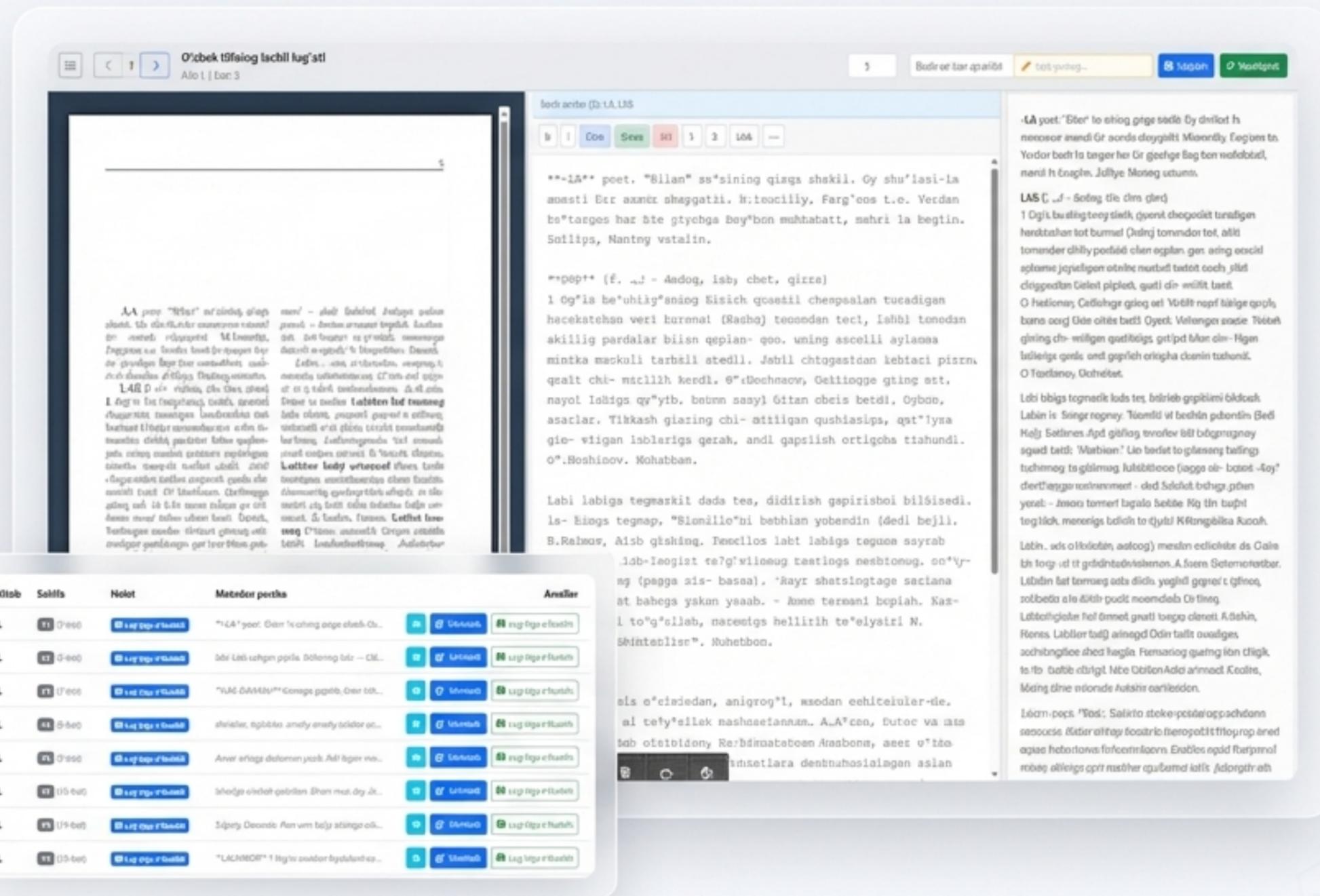
## Step 1: Document Upload

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.

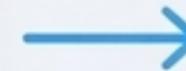
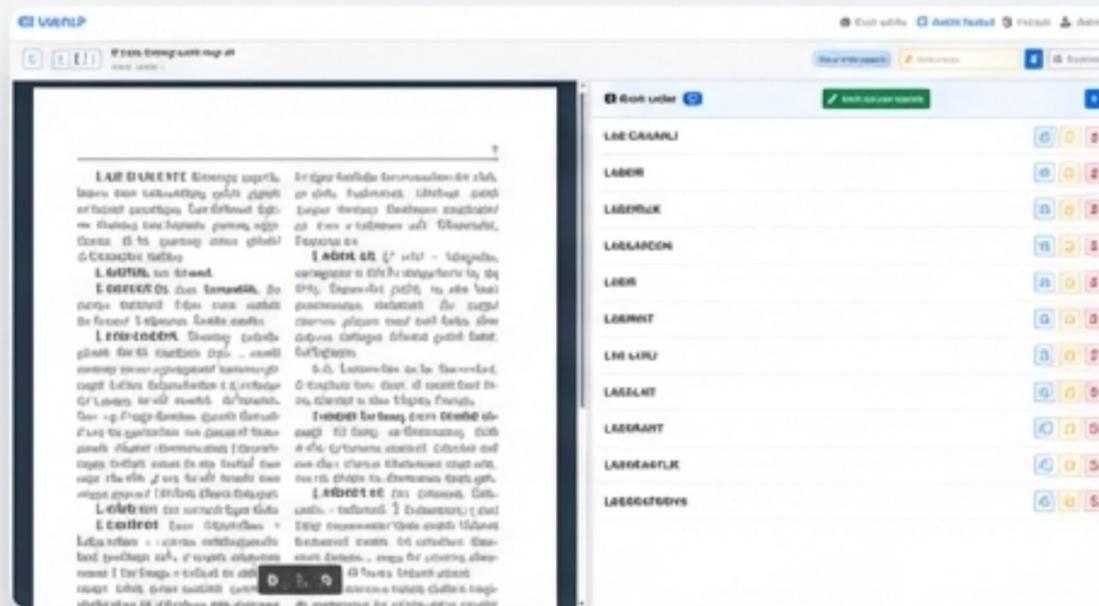
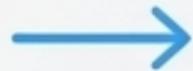
# Pipeline Steps 4, 5 & 6: NLP Parsing to Public API



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

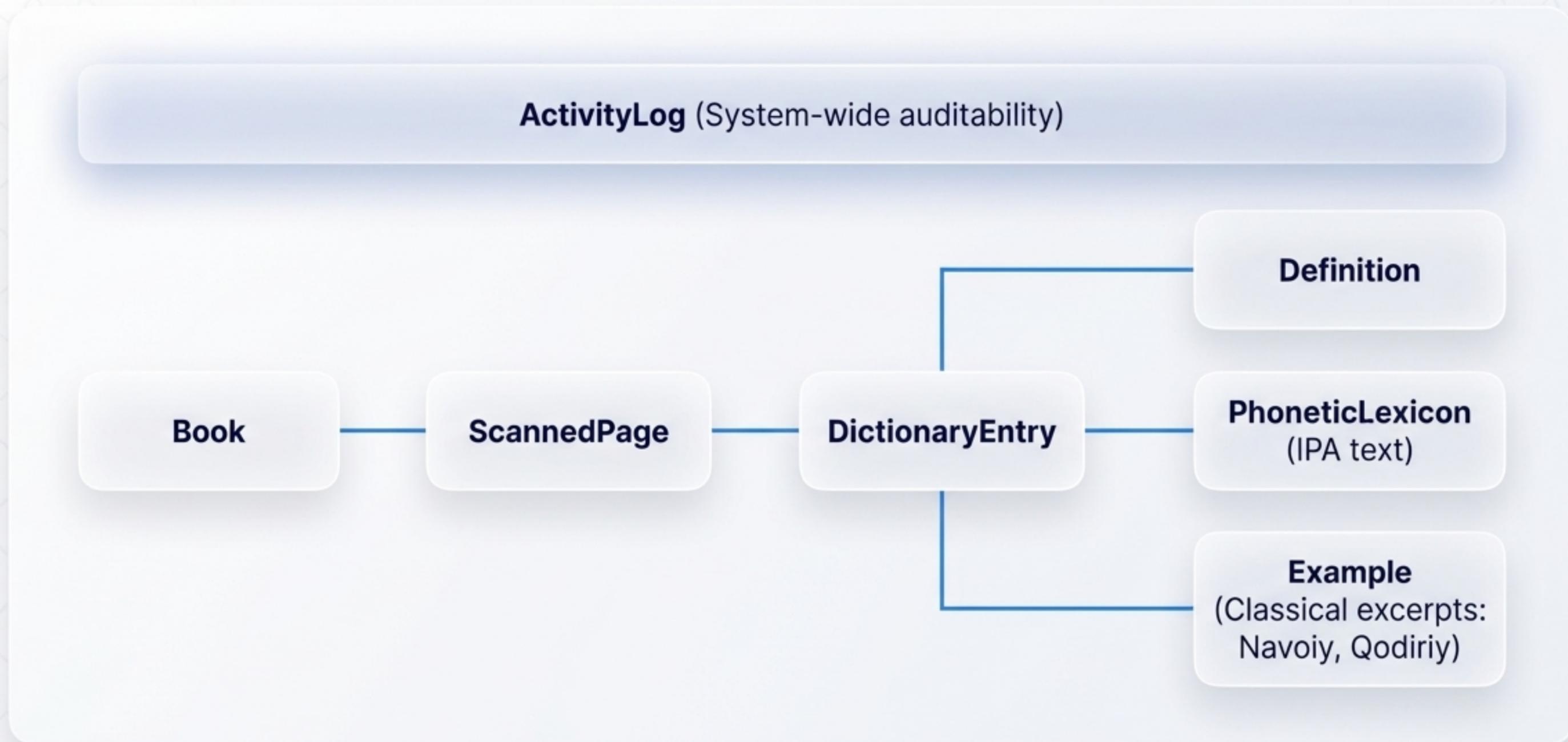
- Secure [Markdown](#) editor
- Contextual verification
- Internal editor notes for printing errors



## 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**

Volumes A, E, G, G',  
H, I, L, N



**Near 100%  
Parsed**

Volumes O, O', F, J



**100% OCR /  
Pending Parsing**

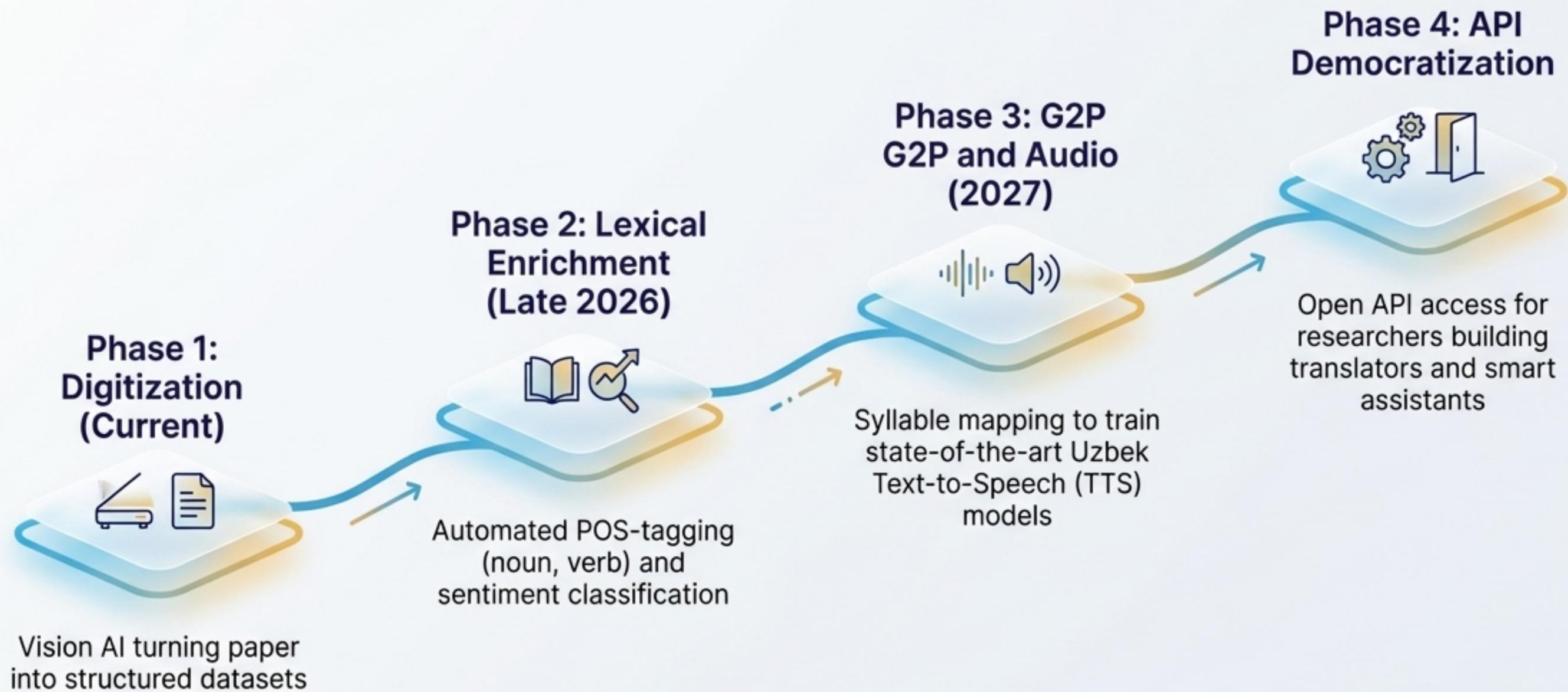
Volumes B, K, M, P, Q,  
R, Sh, Z



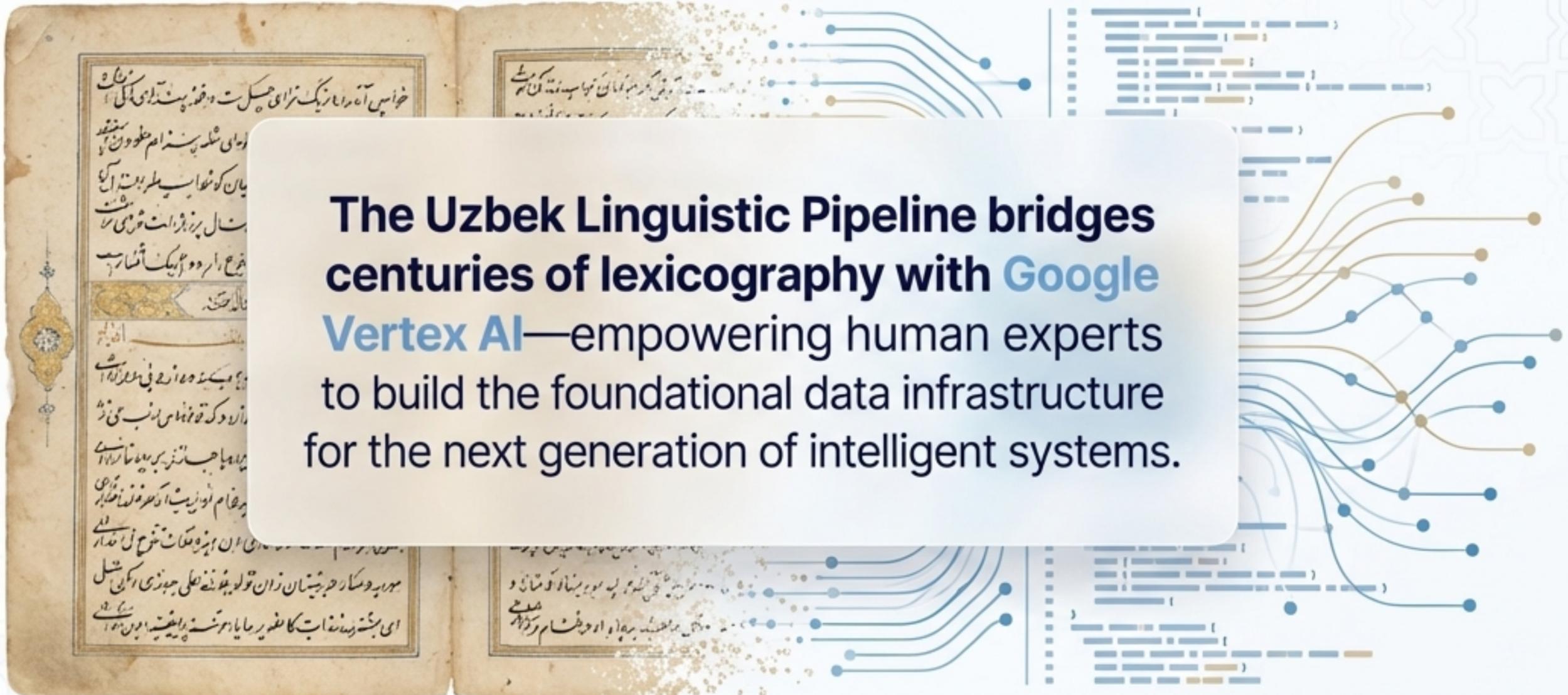
**Currently in  
OCR Queue**

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.**