UniDocVerse: A Privacy-First Offline AI Platform for Universal Document Intelligence

Vijay Kumar Bomma UniDocVerse Technical Report — 2026

Abstract

Organizations across finance, healthcare, legal, and enterprise operations increasingly rely on AI to extract insights from documents. However, existing cloud-based AI tools introduce critical limitations: privacy risks, upload restrictions, limited document format support, and lack of cross-document intelligence. UniDocVerse addresses these challenges through a fully offline, privacy-first, multi-agent document intelligence pipeline capable of processing any document type—PDFs, Word files, spreadsheets, images, emails, and more. This paper presents the UniDocVerse architecture: an 11-agent LangGraph-orchestrated pipeline, adaptive parsing system, hybrid semantic-keyword search engine, pgvector-powered embedding storage, and domain-specific analyzers. We evaluate performance, scalability, and accuracy, and compare UniDocVerse to cloud-based alternatives. Finally, we outline future extensions including email intelligence, calendar extraction, knowledge graph construction, predictive analytics, and autonomous agent workflows.

1. Introduction

Document intelligence has traditionally depended on cloud-based AI systems such as ChatGPT, Copilot, and proprietary OCR platforms. While powerful, these systems introduce four fundamental limitations:

1. Privacy & Compliance Risks

Cloud tools transmit documents to external servers, making them unsuitable for sensitive legal, financial, medical, or government data. Regulations such as HIPAA, GDPR, and SOC2 restrict cloud usage.

2. Limited Document Format Support

Most tools specialize in a single format (e.g., PDFs only). They cannot analyze mixed document collections or generate cross-document insights.

3. Upload & Storage Restrictions

Cloud platforms impose file size limits, pay-per-upload pricing, and lack the ability to process large document repositories.

4. Lack of On-Premise Deployment

Organizations needing air-gapped or offline environments cannot rely on cloud AI.

UniDocVerse solves these challenges by providing a **fully local, multi-agent, universal document intelligence system** capable of processing any document type with no cloud dependency.

2. System Architecture

UniDocVerse is built on a **multi-agent pipeline** consisting of 11 specialized agents orchestrated through LangGraph. Each agent performs a distinct transformation step, enabling a clean, deterministic flow from raw document to structured analytics.

Pipeline Overview

Upload → Ingest → Parse → Cleanup → Doc-Specific Analysis → Analyze → Search → Quality Check → Insights → Metrics → Entity Linking → Finalize → Response

2.1 Agent Descriptions

1. Ingest Agent

Validates file integrity, extracts metadata, and initializes the document state.

2. Parse Agent

Uses the UniversalParser to extract text, tables, and structure from any format. Performs auto-classification and optional specialized re-parsing.

3. Cleanup Agent

Normalizes text, removes NULL bytes, fixes spacing, and preserves metadata.

4. Document-Specific Analysis Agent

Routes documents to specialized analyzers (financial, legal, medical, HR, etc.). Produces domain-specific insights and UI-agnostic visualization schemas.

5. Analyze Agent

Generates summaries and key points using local LLM inference.

6. Search Agent

Creates 768-dim embeddings (all-mpnet-base-v2) and extracts keywords for hybrid search.

7. Quality Check Agent

Scores confidence, detects missing classifications, and identifies parsing issues.

8. Insight Agent

Produces actionable recommendations based on classification and summary.

9. Metrics Agent

Computes word count, character count, and other quantitative metrics.

10. Entity Linking Agent

Extracts entities, builds cross-document relationships, and stores links in PostgreSQL.

11. Finalize Agent

Validates JSON, persists embeddings to pgvector, stores metadata, and returns the final structured response.

3. Technical Implementation

3.1 LangGraph Orchestration

LangGraph was selected for its state-pattern architecture, enabling:

- Shared document state across agents
- Clean separation of concerns
- Conditional routing
- Error handling and retries
- Compatibility with local LLMs (Ollama)

3.2 Vector Storage with pgvector

pgvector enables fully offline semantic search with:

- Zero data leakage
- Unified storage of embeddings + metadata
- Native SQL querying
- Air-gapped deployment

Use cases include similarity search, related document discovery, and entity-based retrieval.

3.3 Embedding Model: all-mpnet-base-v2

Chosen for:

- 768-dimensional vectors (optimal balance of quality and storage)

- Strong performance on semantic similarity tasks
- Fully local inference
- No external API calls

3.4 Adaptive Document Parsing

UniDocVerse uses classification-driven parser selection:

If known type → Specialized Parser + Domain Analyzer

- Bank statements → Financial parser
- Medical records → Healthcare parser
- Invoices → Business parser

If unknown type → Universal Parser

Extracts text, tables, structure, and produces generic insights.

Libraries include PyMuPDF, python-docx, openpyxl, python-pptx, Pillow, pytesseract, and custom parsers.

4. Results & Evaluation

4.1 Processing Performance

Average processing time: **~60 seconds per document**, justified by:

- Full local LLM inference
- 11 sequential agents
- Embedding generation
- Entity linking
- Visualization schema creation

4.2 Batch Processing

Concurrency: **5 documents per batch**, balancing throughput and memory usage.

Documents	Batches	Time
5	1	~1 min
10	2	~2 min
50	10	~10 min
100	20	~20 min

4.3 Storage & Analytics

Each document stores:

- Original file
- Extracted text
- Metadata
- Embeddings
- Insights
- Entity links
- Visualization schemas

4.4 Hybrid Search

UniDocVerse merges:

- **Semantic search** (pgvector cosine similarity)
- **Keyword search** (exact/fuzzy matching)

The system also supports **Try V2**, a re-analysis mechanism for improved insights.

4.5 Comparison with Alternatives

Feature	UniDocVerse	ChatGPT/Copilot	Cloud Tools
Privacy	100% Local	Cloud	Cloud
Internet Required	No	Yes	Yes
Upload Limits	Unlimited	5–10GB	Limited
Multi-format	Universal	Limited	Varies
Cross-doc Linking	Yes	No	Rare
Semantic Search	Local	No	Limited
Data Ownership	Full	None	None
Compliance	HIPAA/GDPR	Risky	Varies
Re-analysis	Yes	No	No

5. Future Work

5.1 Email Intelligence

- Email parsing
- Thread summarization
- Attachment analysis
- Fraud detection
- Cross-email analytics

5.2 Calendar Intelligence

- Appointment extraction
- Calendar sync
- Smart reminders
- Document-linked events

5.3 News Intelligence Engine

- Topic clustering
- Sentiment trends
- Entity tracking
- Risk alerts

5.4 Multi-Domain Data Integration

- Finance, legal, healthcare, HR, real estate
- Cross-domain correlations

5.5 Financial Intelligence Dashboard

- Portfolio tracking
- Earnings call summaries
- Market sentiment

5.6 Knowledge Graph Engine

- Semantic search
- Entity linking
- Trend detection

5.7 Autonomous AI Agent Layer

- Multi-step workflows
- Goal-driven tasks
- Automated insights

5.8 Predictive Intelligence

- Bill forecasting
- Spending prediction
- Workload forecasting
- Document need prediction

6. Conclusion

UniDocVerse introduces a fully offline, privacy-first, multi-agent document intelligence system capable of processing any document type with no cloud dependency. Through LangGraph orchestration, adaptive parsing, hybrid search, and pgvector-powered embeddings, UniDocVerse delivers enterprise-grade analytics, cross-document intelligence, and compliance-ready processing. Future extensions—including email intelligence, calendar extraction, knowledge graphs, and autonomous agents—position UniDocVerse as a universal intelligence engine for organizations of all sizes.