

---

# CONTEXTWEAVER

## Advanced Multi-Document Reasoning Engine





**INFO 7375 - Final Project**



**Northeastern University | Fall 2025**


**Student:** Sravan Kumar Kurapati

**Submission Date:** December 14, 2025

---

 <b>CONTEXTWEAVER</b>	<b>1</b>
Advanced Multi-Document Reasoning Engine	1
 <b>GITHUB REPOSITORY LINKS</b>	<b>5</b>
Main Repository	5
Quick Navigation to Key Sections	5
 <b>DOCUMENTATION</b>	<b>6</b>
GitHub Link	6
Overview	6
Documentation Files	6
1. README.md - Main Project Documentation	6
2. SETUP.md - Installation & Configuration Guide	6
3. USAGE.md - API Reference & Usage Guide	7
4. TESTING.md - Test Results & Validation	8
5. ARCHITECTURE.md - System Design Documentation	9
6. EXAMPLES.md - Example Outputs & Use Cases	9
 <b>SETUP INSTRUCTIONS</b>	<b>11</b>
GitHub Link	11
Overview	11
Installation Summary	11
Prerequisites	11
Installation Steps (5 Steps)	11
Platform-Specific Notes	12
macOS	12
Linux (Ubuntu/Debian)	12
Windows	13
Configuration Details	13
Required Environment Variables	13

Configuration File Location	13
Troubleshooting Guide	13
Issue 1: tiktoken Installation Failed	13
Issue 2: ChromaDB Dimension Mismatch	14
Issue 3: OpenAI API Rate Limit	14
Issue 4: Import Errors	14
Issue 5: Memory Error	14
Issue 6: Streamlit Not Loading	14
Verification Checklist	15
Installation Time Estimate	15
Post-Installation Next Steps	16
 <b>TESTING SCRIPTS</b>	<b>16</b>
GitHub Link	16
Overview	16
Test Suite Structure	16
Main Test File	16
Test Coverage	17
How to Run Tests	18
Run Complete Test Suite	18
Run Individual Component Tests	19
Test Results Summary	19
Verified Metrics (December 12, 2025)	19
Test Data	20
Sample Documents Used	20
Test Outputs Generated	20
Files Created During Testing	20
Continuous Testing	21
When to Run Tests	21
Updating Expected Results	23
Test Quality Metrics	23
Key Testing Achievements	24
 <b>EXAMPLE OUTPUTS</b>	<b>25</b>
GitHub Links	25
Overview	25
Example Files Provided	25
1. example_query_coffee.json	25
2. example_query_chicken.json	26
3. sample_outputs.md	27
Detailed Example Outputs	27
Example 1: Full Pipeline Output	27
Example 2: Multi-Hop Reasoning Chain	30

Example 3: Contradiction Report	31
Example 4: Uncertainty Breakdown	32
Example 5: Synthetic Q&A Pairs Generated	33
Output Formats Available	34
1. JSON Format	35
2. Text Report Format	35
3. Streamlit UI Display	35
Example Output Statistics	35
From Test Run (December 12, 2025)	35
Example Use Case Demonstrations	36
Use Case 1: Research Literature Review	36
Use Case 2: Out-of-Domain Queries	36
Use Case 3: Conflicting Information	36
Accessing Example Outputs	36
In GitHub Repository	36
Generating Your Own Examples	37
Example Quality Metrics	37
 <b>FINE-TUNING DATASETS</b>	<b>38</b>
GitHub Links	38
Overview	38
Why Synthetic Data Instead of Fine-Tuning?	38
Strategic Decision	38
Synthetic Data Generator	39
Implementation	39
What It Does	39
Sample Dataset Provided	39
File: sample_synthetic_dataset.json	39
Dataset Statistics	40
Sample Q&A Pairs	41
Example 1: Easy Difficulty	41
Example 2: Medium Difficulty	41
Example 3: Hard Difficulty	42
Generation Process	42
How Synthetic Data Is Created	42
Use Cases for Synthetic Data	43
1. System Evaluation	43
2. Quality Assurance	43
3. Training Data (Future)	43
4. Benchmarking	44
Comparison: Synthetic Data vs Fine-Tuning	44
Quality Validation	45









Automated Quality Checks	45
Diversity Validation	45
Ethical Validation	45
Scalability	46
Current Sample	46
Production Capability	46
Generation Command	46
Key Achievements	47
 <b>KNOWLEDGE BASE (RAG)</b>	<b>47</b>
GitHub Links	47
Overview	47
Knowledge Base Contents	47
Sample Documents (3 Files)	47
Knowledge Base Design Rationale	48
Why These Specific Documents?	48
Knowledge Base Statistics	49
Overall Metrics	49
Metadata Distribution	50
RAG Component Implementation	50
Requirement Checklist	50
Knowledge Base Organization	52
Hierarchical Structure	53
Automatic Metadata Extraction	53
Knowledge Base Usage	54
How Documents Are Loaded	54
How Documents Are Retrieved	54
Expanding the Knowledge Base	54
How to Add New Documents	54
Supported File Formats	55
Knowledge Base Quality	55
Current Quality Metrics	55
RAG Component Achievements	56

# GITHUB REPOSITORY LINKS

## Main Repository

 [https://github.com/sravankumarkurapati/INFO\\_7375/tree/main/contextweaver](https://github.com/sravankumarkurapati/INFO_7375/tree/main/contextweaver)

## Quick Navigation to Key Sections

Section	GitHub Link	Description
 <b>Complete Source Code</b>	<a href="/src">/src</a>	All 11 Python modules
 <b>Documentation</b>	<a href="/docs">/docs</a>	Setup, Usage, Testing, Architecture, Examples
 <b>Setup Instructions</b>	<a href="docs/SETUP.md">docs/SETUP.md</a>	Detailed installation guide
 <b>Testing Scripts</b>	<a href="tests/test_all_components.py">tests/test_all_components.py</a>	Comprehensive test suite
 <b>Example Outputs</b>	<a href="/examples">/examples</a>	Sample queries and results
 <b>Synthetic Data Sample</b>	<a href="data/synthetic_data">data/synthetic_data</a>	Sample synthetic dataset
 <b>Knowledge Base</b>	<a href="data/sample_docs">data/sample_docs</a>	RAG knowledge base documents
 <b>Streamlit App</b>	<a href="app/streamlit_app.py">app/streamlit_app.py</a>	Interactive demo

---

# DOCUMENTATION

## GitHub Link

 [/docs](#)

---

## Overview

Complete documentation suite covering installation, usage, testing, architecture, and examples. All documentation is written in Markdown format for easy reading on GitHub.

---

## Documentation Files

### 1. README.md - Main Project Documentation

 [README.md](#)

**Purpose:** First point of contact for anyone visiting the repository

#### Contents:

- Project overview and unique value proposition
- Key features summary (3 core components + 4 innovations)
- Quick start guide (web interface + Python API)
- Architecture diagram
- Performance metrics summary
- Technology stack
- Project structure overview
- Contact information

**Length:** ~300 lines

**Audience:** General users, evaluators, potential contributors

---

### 2. SETUP.md - Installation & Configuration Guide

 [docs/SETUP.md](#)

**Purpose:** Complete installation instructions for all platforms

**Contents:**

- **System Requirements** - Python version, RAM, disk space, dependencies
- **Installation Methods** - Standard installation (step-by-step), Docker installation
- **Environment Variables** - All configurable settings with descriptions
- **Platform-Specific Instructions** - macOS, Linux (Ubuntu/Debian), Windows
- **Troubleshooting** - 6 common issues with solutions:
  - tiktoken compilation errors (needs Rust)
  - ChromaDB dimension mismatch
  - OpenAI API rate limits
  - Memory errors
  - LangChain import errors
  - Streamlit not found
- **Verification Checklist** - 9-step checklist to confirm installation
- **Quick Start Commands** - Development and production mode

**Length:** ~400 lines

**Audience:** New users setting up the system

**Key Feature:** Platform-specific troubleshooting for macOS, Linux, Windows

---

### 3. USAGE.md - API Reference & Usage Guide

 [docs/USAGE.md](#)

**Purpose:** Complete API documentation for developers

**Contents:**

- **Quick Start Examples** - Streamlit UI and Python API
- **Core API Reference** - ContextWeaverPipeline class methods
  - `ingest_documents()` - Parameters, return values, supported formats
  - `query()` - All parameters, return structure, options
- **Component Usage** - Individual module usage:
  - Document Processor (4 chunking strategies)
  - Vector Store (similarity search, filtering)
  - Multi-Hop Reasoning (hop configuration)
  - Contradiction Detection (severity levels)
  - Uncertainty Quantification (confidence breakdown)
  - Fact-Checking (claim verification)
  - Knowledge Graph (PageRank, traversal)
  - Hybrid Retrieval (3-tier system)

- **Advanced Features** - Custom prompts, synthetic generation, data augmentation
- **Configuration** - Runtime config, custom weights, model selection
- **Best Practices** - 5 recommendations for optimal usage
- **Common Patterns** - Batch processing, confidence-based decisions, progressive enhancement
- **API Reference Table** - Quick lookup for all classes and methods

**Length:** ~600 lines

**Audience:** Developers integrating ContextWeaver

**Key Feature:** Complete method signatures with parameters and return types

---

## 4. TESTING.md - Test Results & Validation

 [docs/TESTING.md](#)

**Purpose:** Document test suite and verified performance metrics

**Contents:**

- **Test Suite Overview** - 12 comprehensive tests, 100% pass rate
- **Latest Test Results** - December 12, 2025 run
  - Summary: 12/12 passed, 136.68s total time
  - Individual test details with actual metrics
  - Performance grades (A+, A, A-, B+, B)
- **Detailed Results for Each Test:**
  - Test 1: Configuration (PASS)
  - Test 2: Document Processing (PASS - 0.01s)
  - Test 3: Vector Store (PASS - 77.6% similarity)
  - Test 4: Multi-Factor Ranking (PASS - 0.643 score)
  - Test 5: Multi-Hop Reasoning (PASS - 2 hops, 85% confidence)
  - Test 6: Contradiction Detection (PASS - 95% confidence)
  - Test 7: Uncertainty (PASS - 53.4% confidence)
  - Test 8: Fact-Checking (PASS - 100% verified)
  - Test 9: Knowledge Graph (PASS - PageRank working)
  - Test 10: Hybrid Retrieval (PASS - Local 90%, Web 75%)
  - Test 11: Full Pipeline (PASS - 27.7s)
  - Test 12: Synthetic Data (PASS - Quality 94.4%, Diversity 81.9%)
- **Performance Benchmarks** - Speed ratings, accuracy scores
- **Running Tests** - Commands to execute tests
- **Test Data Description** - 3 sample documents used
- **Troubleshooting** - Common test failures and solutions

**Length:** ~550 lines

**Audience:** Technical evaluators, QA testers

**Key Feature:** Actual verified test results with timestamps and metrics

---

## 5. ARCHITECTURE.md - System Design Documentation

 [docs/ARCHITECTURE.md](#)

**Purpose:** Explain system design and technical decisions

**Contents:**

- **System Overview** - 4-layer architecture (Presentation, Orchestration, Processing, Data)
- **Component Architecture** - Detailed breakdown of all 11 modules:
  - Document Processing Pipeline (chunking flow)
  - Vector Storage & Retrieval (embedding flow)
  - Hybrid Retrieval System (3-tier decision tree)
  - Multi-Hop Reasoning (hop algorithm)
  - Prompt Engineering (template system)
  - Uncertainty Quantification (Bayesian formula)
  - Knowledge Graph (graph structure, algorithms)
  - Fact-Checking (verification pipeline)
  - Synthetic Data (generation pipeline)
- **Data Flow** - Complete query processing from input to output
- **Module Details** - Purpose, key classes, performance for each module
- **Design Decisions** - Why hybrid retrieval? Why multi-factor ranking? Why Bayesian uncertainty?
- **Scalability Considerations** - Current capacity, scaling strategies, production optimizations
- **Security & Privacy** - API key protection, data privacy, rate limiting

**Length:** ~800 lines

**Audience:** Technical reviewers, system designers

**Key Feature:** Explains WHY decisions were made, not just WHAT was built

---

## 6. EXAMPLES.md - Example Outputs & Use Cases

 [docs/EXAMPLES.md](#)

**Purpose:** Demonstrate system capabilities with real examples

**Contents:**

- **Example 1: Medical Research Query (Full Pipeline)**
  - Input: "Is moderate coffee safe?"
  - All 8 processing steps shown
  - Actual output from test run
  - Reasoning chain details
  - Full JSON output
- **Example 2: Out-of-Domain Query (Web Fallback)**
  - Input: "Is chicken healthy?"
  - Demonstrates automatic web fallback
  - Shows difference in confidence (75% vs 90%)
  - Simplified processing for web sources
- **Example 3: Contradiction Detection**
  - Input: 2 conflicting documents
  - Detection output with severity (HIGH)
  - Explanation of why they contradict
  - Resolution strategy
- **Example 4: Multi-Hop Reasoning**
  - Query requiring 2 reasoning steps
  - Hop-by-hop breakdown
  - Citation tracking
  - Final synthesis
- **Example 5: Synthetic Data Generation**
  - 3 generated Q&A pairs (easy, medium, hard)
  - Quality assessment (94.4%)
  - Diversity metrics (81.9%)
- **Performance Examples** - Query times, token usage, costs
- **Visualization Examples** - Graph visualization, uncertainty gauge
- **Code Examples** - Basic usage, advanced usage, component-level usage
- **Real-World Use Cases** - Legal analysis, research review scenarios

**Length:** ~700 lines

**Audience:** Users wanting to see what system can do

**Key Feature:** Real outputs from actual test runs, not hypothetical

# SETUP INSTRUCTIONS

## GitHub Link

 [docs/SETUP.md](#)

---

## Overview

Complete step-by-step installation and configuration guide for ContextWeaver. Covers all platforms (macOS, Linux, Windows) with troubleshooting for common issues.

---

## Installation Summary

### Prerequisites

Requirement	Minimum	Recommended
Python	3.9+	3.10+
RAM	4GB	8GB
Disk Space	2GB	5GB
Internet	Required	Stable connection
OpenAI API	Account with credits	\$10+ balance

### Installation Steps (5 Steps)

#### Step 1: Clone Repository

```
bash
git clone https://github.com/YOUR_USERNAME/contextweaver.git
cd contextweaver
```

#### Step 2: Create Virtual Environment

```
bash
python3 -m venv venv
source venv/bin/activate # macOS/Linux
# or venv\Scripts\activate on Windows
```

### Step 3: Install Dependencies

```
bash
pip install --upgrade pip
pip install -r requirements.txt
```

### Step 4: Configure Environment

```
bash
cp .env.example .env
# Edit .env and add your OPENAI_API_KEY
```

### Step 5: Verify Installation

```
bash
python -c "from src.config import Config; Config.validate(); print('✅ Ready!)"
```

**Total Time:** ~5-10 minutes

---

## Platform-Specific Notes

### macOS

- **Rust Compiler Required** - For tiktoken package

```
bash
brew install rust
```

- **M1/M2 Chips** - Some packages may need Rosetta 2
- **Python Installation** - Use Homebrew: `brew install python@3.10`

### Linux (Ubuntu/Debian)

- **Build Tools Required** - For compiling packages

bash

`sudo apt install build-essential cargo python3-dev`

- **Python Version** - Ensure 3.9+ is installed

## Windows

- **Visual C++ Build Tools** - May be needed for some packages
  - **Python PATH** - Ensure Python is added to PATH during installation
  - **PowerShell vs CMD** - Commands may differ slightly
- 

## Configuration Details

### Required Environment Variables

Must Set:

- **OPENAI\_API\_KEY** - Your OpenAI API key (starts with `sk-`)

Optional (Have Defaults):

- **MODEL\_NAME** - Default: `gpt-4-turbo-preview`
- **EMBEDDING\_MODEL** - Default: `text-embedding-3-small`
- **CHUNK\_SIZE** - Default: `1000`
- **TOP\_K** - Default: `10`
- **LOG\_LEVEL** - Default: `INFO`

### Configuration File Location

- File: `.env` in project root
  - Template: `.env.example` (copy this)
  - **Never commit** `.env` - Contains API keys!
- 

## Troubleshooting Guide

### Issue 1: tiktoken Installation Failed

**Symptom:** Error about missing Rust compiler

**Solution:**

bash

*# macOS: brew install rust*

*# Ubuntu: sudo apt install cargo*

*# Windows: Download from [rust-lang.org](https://rust-lang.org)*

pip install --force-reinstall tiktoken

## Issue 2: ChromaDB Dimension Mismatch

**Symptom:** InvalidDimensionException

**Solution:**

bash

rm -rf data/chroma\_db/\*

*# Restart app (will recreate with correct dimensions)*

## Issue 3: OpenAI API Rate Limit

**Symptom:** RateLimitError from OpenAI

**Solution:** Wait 1 minute or upgrade OpenAI plan

## Issue 4: Import Errors

**Symptom:** "No module named 'langchain'"

**Solution:**

bash

pip install -r requirements.txt --force-reinstall

## Issue 5: Memory Error

**Symptom:** Cannot allocate memory

**Solution:** Reduce CHUNK\_SIZE and TOP\_K in .env

## Issue 6: Streamlit Not Loading

**Symptom:** Streamlit command not found

**Solution:**

bash

pip install streamlit==1.30.0

```
streamlit run app/streamlit_app.py
```

---

## Verification Checklist

After installation, verify:

- Virtual environment activated
- All packages installed (check with `pip list`)
- `.env` file created with API key
- Config validation passes
- Sample documents exist in `data/sample_docs/`
- Can import modules: `from src.config import Config`
- Streamlit launches: `streamlit run app/streamlit_app.py`
- Test query works
- No import errors

### Run Verification Script:

```
bash
```

```
python -c "from src.config import Config; Config.validate(); print('✅ All checks passed!)"
```


---

## Installation Time Estimate

Task	Time
Clone repository	30 seconds
Create virtual environment	1 minute
Install dependencies	3-5 minutes
Configure <code>.env</code>	1 minute
Verify installation	30 seconds
<b>Total</b>	<b>5-10 minutes</b>

---

## Post-Installation Next Steps

1.  Installation complete
2. → Read [USAGE.md](#) for API reference
3. → Try example queries in Streamlit app
4. → Run test suite: `python tests/test_all_components.py`
5. → Explore [EXAMPLES.md](#) for use cases

## TESTING SCRIPTS

### GitHub Link

 [tests/test\\_all\\_components.py](#)

---

### Overview

Comprehensive test suite that validates all 12 components of ContextWeaver with **100% test pass rate** verified on December 12, 2025.

---

### Test Suite Structure

#### Main Test File

**File:** `tests/test_all_components.py`

**Lines of Code:** 590 lines

**Total Tests:** 12 comprehensive tests

**Execution Time:** 136.68 seconds (~2.3 minutes)

## Test Coverage

Test #	Component Tested	What It Validates	Status	Time
1	Configuration	All settings, API keys, paths valid	✓ PASS	<0.01 s
2	Document Processing	File loading, chunking (4 strategies), metadata extraction	✓ PASS	0.01s
3	Vector Store	Embeddings creation, ChromaDB storage, similarity search	✓ PASS	5.70s
4	Multi-Factor Ranking	5-factor ranking algorithm (similarity + metadata)	✓ PASS	<1s
5	Multi-Hop Reasoning	Cross-document reasoning, citation tracking	✓ PASS	25.2s
6	Contradiction Detection	Finding conflicting claims, severity classification	✓ PASS	9.0s
7	Uncertainty Quantification	Bayesian confidence, component scoring	✓ PASS	<0.01 s
8	Fact-Checking	Claim verification, source cross-checking	✓ PASS	1.0s

9	Knowledge Graph	Graph construction, PageRank calculation	✓ PASS	<0.01 s
10	Hybrid Retrieval	3-tier fallback (Local→Web→LLM)	✓ PASS	<5s
11	Full Pipeline	Complete end-to-end integration	✓ PASS	27.7s
12	Synthetic Data	Q&A generation, quality/diversity metrics	✓ PASS	27.7s

**Success Rate:** 12/12 (100%) 🎉

---

# How to Run Tests

## Run Complete Test Suite

bash

*# Activate virtual environment*

*source venv/bin/activate*

*# Run all 12 tests*

*python tests/test\_all\_components.py*

*# Expected output:*

*# ✓ Passed: 12/12 (100%)*

*# 🕒 Total Time: ~136s*

#  Results saved to: test\_outputs/comprehensive\_test\_results.json

## Run Individual Component Tests

Each module has built-in tests:

bash

```
python src/document_processor.py    # Test chunking
```

```
python src/vector_store.py          # Test embeddings
```

```
python src/reasoning_engine.py      # Test multi-hop
```

```
python src/uncertainty_quantification.py # Test Bayesian
```

---

## Test Results Summary

### Verified Metrics (December 12, 2025)


#### Performance Metrics:

- Document Processing Speed: **0.01s** for 3 files ⚡
- Vector Embedding Speed: **5.70s** for 3 documents
- Similarity Search Accuracy: **77.6%** match score
- Multi-Hop Reasoning: **2 hops** used, **85% confidence**
- Contradiction Detection: **95% confidence**, **9.0s** processing
- Fact-Checking Accuracy: **100% verified** (HIGHLY VERIFIED)
- Synthetic Data Quality: **94.4%** (Outstanding)
- Synthetic Data Diversity: **81.9%** (Excellent)

#### Accuracy Metrics:

- Hybrid Retrieval (Local): **90% confidence** ★
- Hybrid Retrieval (Web): **75% confidence** ★
- Overall Confidence: **69.9%** (MODERATE - appropriately calibrated)
- PageRank Accuracy: **48.1%** for most important doc (2023 study)

#### Integration Test:

- All 11 modules working together:  **Success**
- Full pipeline execution: **27.7 seconds**
- No component conflicts or errors

- Coherent end-to-end output
- 

## Test Data

### Sample Documents Used

#### File 1: coffee\_study\_2018.txt

- Purpose: Baseline study with methodological limitations
- Year: 2018
- Finding: High coffee consumption increases risk
- Use: Tests contradiction detection





#### File 2: coffee\_study\_2023.txt

- Purpose: Recent study with improved methodology
- Year: 2023
- Finding: Moderate coffee consumption is protective
- Use: Tests recency ranking, resolution

#### File 3: meta\_analysis\_2022.txt

- Purpose: Meta-analysis explaining contradictions
- Year: 2022
- Finding: Confounders explain earlier negative findings
- Use: Tests synthesis, temporal analysis

#### Why These Documents:

-  Creates temporal range (6 years)
  -  Contains contradictions (for detection testing)
  -  Shows knowledge evolution (for multi-hop)
  -  Different credibility levels (for ranking)
- 

## Test Outputs Generated

### Files Created During Testing

1. comprehensive\_test\_results.json

- Location: `test_outputs/comprehensive_test_results.json`
- Size: ~5 KB
- Contents: Complete test results in JSON format
- Includes: Metadata, all 12 test results, summary statistics

## 2. Console Output

- Real-time progress for each test
- Color-coded status (✓ PASS, ✗ FAIL)
- Performance metrics displayed
- Component-level details

## 3. Log Files (if debug enabled)

- Location: `logs/` directory
  - Contents: Detailed execution logs
  - Useful for debugging
- 

# Continuous Testing

## When to Run Tests

### Before Committing Code:

bash

```
python tests/test_all_components.py
```

*# Ensure 12/12 pass before git commit*

### After Changing Configuration:

bash

```
python tests/test_all_components.py
```

*# Verify config changes don't break system*

### After Adding New Documents:

bash

*# Test with new documents*

```
python -c "  
  
from src.contextweaver_pipeline import ContextWeaverPipeline  
  
pipeline = ContextWeaverPipeline()  
  
pipeline.ingest_documents(['your_new_doc.pdf'])  
  
# Check for errors  
  
"
```

### **Before Production Deployment:**

```
bash  
  
# Full validation  
  
python tests/test_all_components.py  
  
streamlit run app/streamlit_app.py # Manual UI testing  
  
...
```

---

*## Test Maintenance*

*### Adding New Tests*

Tests follow this structure:

```
...  
  
def test_N_new_feature():  
  
    print_section("TEST N: New Feature")  
  
    try:
```

```
# Setup

# Execute test

# Validate results

return {"status": "PASS", "metrics": {...}}

except Exception as e:

    return {"status": "FAIL", "error": str(e)}
```

## Updating Expected Results

When system improves, update expected values:

- Similarity thresholds
  - Performance benchmarks
  - Quality scores
- 

## Test Quality Metrics

### Code Quality:

- ☒ Comprehensive coverage (100% of components)
- ☒ Clear test structure
- ☒ Detailed output messages
- ☒ JSON result export
- ☒ Error handling with stack traces

### Reliability:






- ☒ 100% pass rate on verified run
- ☒ Deterministic results (same input → same output)
- ☒ No flaky tests
- ☒ Clear success/failure criteria

### Maintainability:

- ☒ Modular test functions
- ☒ Reusable helper functions
- ☒ Well-documented
- ☒ Easy to extend

---

## Key Testing Achievements

-  **100% Test Pass Rate** - All 12 tests passing
-  **Fast Execution** - Complete suite in 2.3 minutes
-  **Comprehensive Coverage** - Every component validated
-  **Real Metrics** - Actual performance measured, not estimated
-  **Repeatable** - Consistent results across runs



# EXAMPLE OUTPUTS

## GitHub Links

[/examples](#)  
[docs/EXAMPLES.md](#)

---

## Overview

Real example outputs from actual test runs demonstrating ContextWeaver's capabilities. All examples verified on December 12, 2025.

---

## Example Files Provided

### 1. example\_query\_coffee.json

[examples/example\\_query\\_coffee.json](#)

**Query:** "Is moderate coffee consumption safe for heart health?"

#### Demonstrates:

- Local knowledge base retrieval (90% confidence)
- Multi-hop reasoning (1 hop)
- Contradiction detection (2 found)
- Uncertainty quantification (69.9% confidence)
- Fact-checking (33% verified)
- Complete JSON output structure

#### Key Results:

Source: LOCAL (from knowledge base)

Confidence: 69.9% (MODERATE)

Hops: 1 reasoning step

Contradictions: 2 detected (HIGH severity)

Processing Time: 27.69s

Answer: "Yes, moderate coffee consumption is considered safe..."

---

## 2. example\_query\_chicken.json

 [examples/example\\_query\\_chicken.json](#)

**Query:** "Is chicken healthy?"

**Demonstrates:**

- Web search fallback (75% confidence)
- Out-of-domain query handling
- Simplified processing for web sources
- Graceful degradation

**Key Results:**

Source: WEB (fallback tier 2)

Confidence: 75% (MEDIUM)

Fallback Reason: "Local documents have low relevance to query"

Web Results: 5 sources

Processing Time: ~10s

Answer: "Yes, chicken is a healthy protein source..."

**Comparison with Coffee Query:**

Metric	Coffee (Local)	Chicken (Web)

Source	LOCAL KB	WEB SEARCH
Confidence	90%	75%
Processing	Full pipeline	Simplified
Components	All 6 enabled	Basic only
Time	27.7s	~10s

---

### 3. sample\_outputs.md

 [examples/sample\\_outputs.md](#)

**Purpose:** Quick reference showing typical outputs for different query types

**Contains:**

- **Query 1:** Coffee safety (local KB) - Full output with all metrics
- **Query 2:** Chicken health (web fallback) - Shows tier 2 activation
- **Query 3:** Evolution analysis (temporal) - Shows temporal reasoning

**Format:** Markdown with formatted output examples

---

## Detailed Example Outputs

### Example 1: Full Pipeline Output

**Input Query:**

"Is moderate coffee consumption safe for heart health?"

### Complete Output Structure:

```
{  
  "query": "Is moderate coffee consumption safe for heart health?",  
  "answer": "Yes, moderate coffee consumption is considered safe...",  
  
  "retrieval": {  
    "retrieval_source": "local",  
    "num_documents": 4,  
    "retrieval_confidence": 0.9,  
    "top_sources": [  
      "coffee_study_2023.txt",  
      "meta_analysis_2022.txt",  
      "coffee_study_2018.txt"  
    ]  
  },  
  
  "reasoning": {  
    "hops_used": 1,  
    "confidence": 0.85,  
    "reasoning_chain": [...]  
  },  
  
  "contradictions": {  
    "num_contradictions": 2,
```

```
"overall_severity": "HIGH",  
"contradictions": [...]  
},  
  
"uncertainty": {  
  "confidence_score": 0.699,  
  "confidence_level": "MODERATE",  
  "component_scores": {...}  
},  
  
"fact_check": {  
  "overall_score": 0.33,  
  "verification_level": "PARTIALLY VERIFIED",  
  "num_verified": 2,  
  "num_claims": 6  
},  
  
"performance": {  
  "query_time_seconds": 27.69,  
  "api_calls": 7,  
  "tokens_used": 3200  
}  
}
```

---

## Example 2: Multi-Hop Reasoning Chain

### Hop 1 Output:

Extracted Information:

"2018 study found high coffee consumption increases cardiovascular risk, but 2022 meta-analysis identified methodological flaws - lack of confounder controls."

Connection to Context:

"Early studies had design issues affecting their conclusions."

Intermediate Conclusion:

"Evidence quality questionable due to uncontrolled confounders."

Sufficient Info: NO → Continue to Hop 2

### Hop 2 Output:

Extracted Information:

"2023 research with rigorous controls shows moderate coffee consumption is beneficial for heart health."

Connection to Previous:

"When confounders are controlled, results differ significantly."

Key distinction: MODERATE vs HIGH consumption."

Final Conclusion:

"Moderate coffee consumption (2-3 cups/day) is safe and potentially beneficial based on latest controlled research."

Sufficient Info: YES → Stop reasoning

---

### **Example 3: Contradiction Report**

#### **Detected Contradiction:**

Claim A: "High coffee consumption increases cardiovascular risk by 23%"

Source A: study\_2018.txt (Year: 2018)

Claim B: "Moderate coffee consumption shows 15% reduction in cardiovascular risk"

Source B: study\_2023.txt (Year: 2023)

Severity: HIGH

Confidence: 95%

Explanation:

"The contradiction arises from different consumption levels studied (high: >5 cups vs moderate: 2-3 cups) and improved methodology in later studies that controlled for confounding variables."

Resolution Strategy: TEMPORAL + METHODOLOGICAL

- Trust newer study (2023) over older (2018)
- Recognize different dosage levels

- Acknowledge methodology improvements
- 

## Example 4: Uncertainty Breakdown

### Component Scores:

Evidence Sufficiency: 65.0%

- Based on: 2 supporting sources
- Gap: Recommend 5+ sources for high confidence

Source Agreement: 95.0%

- Domain agreement: 100% (all research)
- Temporal agreement: 90% (within 6 years)

Source Quality: 92.5%

- Average credibility: 0.925
- All medium-to-high quality sources

Contradiction Penalty: 30.1%

- 1 contradiction detected
- Logarithmic penalty applied
- Reduces confidence by ~30%

Final Confidence: 53.4% (MODERATE)

### Sensitivity Analysis:

What-If Scenarios:

Add high-quality source: 57% (+3.6%) ↑

Remove contradictions: 76% (+22.6%) ↑

Add contradiction: 48% (-5.4%) ↓

Double evidence: 61% (+7.6%) ↑

---

## Example 5: Synthetic Q&A Pairs Generated

### Easy Difficulty:

Question: "What did the 2023 study find about moderate coffee consumption?"

Answer: "The 2023 study found that moderate coffee consumption (2-3 cups per day) is associated with a 15% reduction in cardiovascular disease risk when confounding factors are properly controlled."

Difficulty: easy

Requires Docs: 1

Quality Score: 100%

### Medium Difficulty:

Question: "How does the recent research on protective cardiovascular effects of coffee compare to earlier findings?"

Answer: "Recent 2023 research contrasts with earlier 2018 findings.

The 2018 study found increased risk but lacked controls for

confounders. The 2023 study, with rigorous methodology, found protective effects for moderate consumption."

Difficulty: medium

Requires Docs: 2

Reasoning Steps: 3

Quality Score: 95%

**Hard Difficulty:**

Question: "Based on the evolution of coffee research from 2018-2023, what can we conclude about the role of confounding variables in cardiovascular studies?"

Answer: "The evolution demonstrates the critical importance of controlling confounding variables. Early studies showing negative effects failed to account for smoking, sugar, and exercise. When the 2023 study implemented rigorous controls, it found opposite results - protective effects instead of harm."

Difficulty: hard

Requires Docs: 3

Reasoning Steps: 4

Quality Score: 89%

---

**Output Formats Available**

## 1. JSON Format

- Machine-readable
- Complete data structure
- All metrics included
- Easy to parse programmatically

## 2. Text Report Format

- Human-readable
- Formatted sections
- Summary statistics
- Downloadable from Streamlit app

## 3. Streamlit UI Display

- Interactive visualization
  - Real-time metrics
  - Color-coded components
  - Export buttons
- 

# Example Output Statistics

## From Test Run (December 12, 2025)

### Documents Analyzed:

- Total: 3 sample documents
- Chunks Created: 3
- Vector Embeddings: 36 total in database
- Graph Nodes: 3
- Graph Edges: 1-4 (depending on relationships)

### Query Results:

- Queries Tested: 5+
- Average Response Time: 15-28 seconds
- Average Confidence: 60-85%
- Average Documents Retrieved: 4-5

### Synthetic Data Generated:

- Q&A Pairs: 3 samples

- Quality: 94.4%
  - Diversity: 81.9%
  - Generation Time: 27.65s
- 

## Example Use Case Demonstrations

### Use Case 1: Research Literature Review

**Scenario:** Analyzing multiple research papers on same topic

**Demonstrates:** Multi-hop reasoning, temporal analysis, contradiction detection

**Example:** Coffee health research evolution 2018-2023

### Use Case 2: Out-of-Domain Queries

**Scenario:** Query outside knowledge base scope

**Demonstrates:** Hybrid retrieval, web fallback, confidence calibration

**Example:** Chicken health query (not in local KB)

### Use Case 3: Conflicting Information

**Scenario:** Multiple sources with contradictions

**Demonstrates:** Contradiction detection, resolution strategies, uncertainty

**Example:** High vs moderate coffee consumption findings

---

## Accessing Example Outputs

### In GitHub Repository

examples/

|— example\_query\_coffee.json    # Full pipeline example

|— example\_query\_chicken.json    # Web fallback example

|— sample\_outputs.md            # Quick reference

test\_outputs/

└─ comprehensive\_test\_results.json # Complete test results

## Generating Your Own Examples

### Via Streamlit App:

1. Launch: `streamlit run app/streamlit_app.py`
2. Enter query
3. Click "Download JSON" or "Download Report"
4. Results saved to your downloads folder

### Via Python API:

```
python
```

```
pipeline = ContextWeaverPipeline()
```

```
result = pipeline.query("Your question")
```

```
# Save to file
```

```
import json
```








```
with open('my_example.json', 'w') as f:
```

```
    json.dump(result, f, indent=2, default=str)
```

---

## Example Quality Metrics

### All examples include:

-  Complete input/output pairs
-  Actual timestamps
-  Real performance metrics
-  Verified results (not hypothetical)
-  Component-level breakdowns
-  Confidence scores
-  Processing times

**Examples verified:** December 12, 2025, 19:16:15 EST



# FINE-TUNING DATASETS

## GitHub Links

[data/synthetic\\_data/](#)

[sample\\_synthetic\\_dataset.json](#)

---

## Overview

ContextWeaver implements the **Synthetic Data Generation** core component instead of Fine-Tuning. This was a strategic decision based on the 7-day timeline and focus on more impactful innovations.

---


## Why Synthetic Data Instead of Fine-Tuning?

### Strategic Decision

#### Fine-Tuning Requirements:

- Collect 500-1,000+ training examples
- Prepare and format datasets
- Fine-tune model (hours of compute)
- Validate and test fine-tuned model
- **Estimated Time:** 20-30 hours

#### Synthetic Data Benefits:

- Demonstrates data generation capabilities
- Creates evaluation benchmarks
- Supports testing and validation
- Can be used for future fine-tuning
- **Actual Time:** 3-4 hours 

**Decision:** Implement Synthetic Data Generation (Core Component #3) and focus development time on innovations (Hybrid Retrieval, Uncertainty, Fact-Checking, Knowledge Graph).

**Result:** 3 core components implemented (150% of requirement) + 4 major innovations

---

# Synthetic Data Generator

## Implementation

**File:** `src/synthetic_data_generator.py`

**Lines of Code:** 620 lines

**Status:**  Fully implemented and tested

## What It Does

### 1. Q&A Pair Generation

- Generates question-answer pairs from source documents
- 3 difficulty levels: easy, medium, hard
- Includes reasoning steps for complex questions
- Source attribution maintained

### 2. Data Augmentation

- Query paraphrasing (3 variations per query)
- Document variations (paraphrase, simplify, formalize)
- Noise injection for robustness testing

### 3. Quality Assurance

- Quality scoring (94.4% achieved)
- Diversity checking (81.9% achieved)
- Automatic filtering of low-quality samples
- Ethical sanitization (PII removal, bias detection)

### 4. Benchmark Creation

- Single-hop Q&A
- Multi-hop reasoning
- Contradiction detection
- Temporal analysis

---

## Sample Dataset Provided

**File:** `sample_synthetic_dataset.json`

## Contents:

- **3 Q&A Pairs** (1 easy, 1 medium, 1 hard)
- **Metadata** - Generation timestamp, quality scores, diversity metrics
- **Quality Metrics** - 94.4% overall quality
- **Diversity Metrics** - 81.9% overall diversity
- **Ethical Compliance** - No PII, bias score 0.12, no harmful content

## Dataset Statistics

Total Samples: 3 (sample size)

Generation Time: 27.65 seconds

Size: ~3 KB

### Quality Breakdown:

Overall Quality: 94.4%

High-Quality Ratio: 100%

Completeness: 100%

Coherence: 88%

### Diversity Breakdown:

Overall Diversity: 81.9%

Lexical Diversity: 45.7%

Difficulty Distribution: 100% (perfect balance)

### Ethical Compliance:

No PII: 

Bias Score: 0.12 (low bias)

Harmful Content:  None

## Sample Q&A Pairs

### Example 1: Easy Difficulty

**Question:**

"What did the 2023 study find about moderate coffee consumption?"

**Answer:**

"The 2023 study found that moderate coffee consumption (2-3 cups per day) is associated with a 15% reduction in cardiovascular disease risk when confounding factors are properly controlled."

**Metadata:**

- Difficulty: easy
  - Requires Documents: 1
  - Source: coffee\_study\_2023.txt
  - Generated: 2025-12-12T19:16:15
- 

### Example 2: Medium Difficulty

**Question:**

"How does the recent research on the protective cardiovascular effects of coffee consumption compare to earlier findings?"

**Answer:**

"Recent 2023 research contrasts with earlier 2018 findings. The 2018 study found increased risk but lacked controls for confounders. The 2023 study, with rigorous methodology, found protective effects for moderate consumption. The difference is explained by improved methodology and focus on moderate vs high consumption levels."

**Metadata:**

- Difficulty: medium
- Requires Documents: 2
- Reasoning Steps: 3
- Sources: coffee\_study\_2018.txt, coffee\_study\_2023.txt
- Generated: 2025-12-12T19:16:15

---

### Example 3: Hard Difficulty

#### Question:

"Based on the evolution of coffee research from 2018-2023, what can we conclude about the role of confounding variables in cardiovascular studies?"

#### Answer:

"The evolution demonstrates the critical importance of controlling confounding variables. Early studies showing negative effects failed to account for smoking, sugar intake, and exercise. When the 2023 study implemented rigorous controls, it found opposite results - protective effects instead of harm. This illustrates how uncontrolled confounders can completely reverse study conclusions in cardiovascular research."

#### Metadata:

- Difficulty: hard
  - Requires Documents: 3
  - Reasoning Steps: 4
  - Sources: coffee\_study\_2018.txt, meta\_analysis\_2022.txt, coffee\_study\_2023.txt
  - Generated: 2025-12-12T19:16:15
- 

## Generation Process

### How Synthetic Data Is Created

#### Step 1: Source Selection

- Randomly select N documents based on difficulty
- Easy: 1 document
- Medium: 2 documents
- Hard: 3+ documents

#### Step 2: LLM-Based Generation

- Use GPT-4 with specialized prompts
- Generate question relevant to selected documents
- Generate answer with citations and reasoning
- Include difficulty metadata

#### Step 3: Quality Filtering

- Check completeness (question + answer present)
- Validate length (Q: 10-200 chars, A: 20-1,000 chars)
- Verify coherence (proper formatting, ends with punctuation)
- Ensure reasoning steps for hard difficulty
- **Threshold:** 70% minimum quality score
- **Achieved:** 94.4% average quality ★

#### Step 4: Diversity Checking

- Calculate lexical diversity (unique words / total words)
- Check difficulty distribution (uniform across easy/medium/hard)
- Measure length variance
- **Threshold:** 60% minimum diversity
- **Achieved:** 81.9% average diversity ★

#### Step 5: Ethical Sanitization

- Remove PII patterns (SSN, email, phone, names)
  - Detect bias (topic balance, source diversity)
  - Filter harmful content
  - Maintain attribution
- 

## Use Cases for Synthetic Data

### 1. System Evaluation

- Test multi-hop reasoning capabilities
- Validate contradiction detection
- Benchmark against ground truth

### 2. Quality Assurance

- Ensure consistent performance
- Identify edge cases
- Validate answer quality





### 3. Training Data (Future)

- Can be used to fine-tune models
- Demonstrates data preparation capability
- Ready for larger-scale generation

## 4. Benchmarking

- Compare ContextWeaver vs baseline systems
  - Measure improvement over time
  - Validate new features
- 

## Comparison: Synthetic Data vs Fine-Tuning

Aspect	Fine-Tuning	Synthetic Data (Implemented)
Time Required	20-30 hours	3-4 hours 
Complexity	High	Medium
Dataset Size	500-1,000+ samples	Scalable (generated as needed)
Model Changes	Permanent	No model changes
Validation	Complex	Straightforward
Demonstrates Component	Yes (Fine-Tuning)	Yes (Synthetic Data Generation) 
Innovation Potential	Limited	High (quality/diversity metrics) 
Project Fit	Difficult in 7 days	Perfect for timeline 






---

# Quality Validation

## Automated Quality Checks

Implemented in: `synthetic_data_generator.py` → `QualityChecker` class

### Checks Performed:

1.  Completeness (all required fields present)
2.  Length appropriateness (within acceptable ranges)
3.  Coherence (proper sentence structure, punctuation)
4.  Reasoning steps (present for medium/hard difficulty)
5.  Source attribution (documents tracked)




### Results:

- Overall Quality: **94.4%** (Outstanding)
- High-Quality Ratio: **100%** (All samples above threshold)

## Diversity Validation

Implemented in: `synthetic_data_generator.py` → `DiversityChecker` class

### Metrics Calculated:

1.  Lexical diversity (45.7% - unique word ratio)
2.  Difficulty distribution (100% - perfectly balanced)
3.  Length variance (High - good variation)





### Results:

- Overall Diversity: **81.9%** (Excellent)

## Ethical Validation

Implemented in: `synthetic_data_generator.py` → `EthicalConsiderations` class

### Protections Applied:

1.  PII detection and removal (SSN, email, phone patterns)
2.  Bias detection (topic balance, source diversity)
3.  Content safety (no harmful content generated)
4.  Attribution (source provenance maintained)

### Results:

- No PII:  Confirmed
  - Bias Score: **0.12** (Low - acceptable)
  - Harmful Content:  None detected
- 

## Scalability

### Current Sample

- Size: 3 Q&A pairs
- Purpose: Demonstration
- File Size: ~3 KB

### Production Capability

- Can generate: 50-100+ pairs per run
- Configurable difficulty distribution
- Adjustable quality thresholds
- Batch generation supported

### Generation Command

bash

```
python -c "
```

```
from src.synthetic_data_generator import SyntheticDataGenerator
```

```
from src.document_processor import DocumentProcessor
```

```
processor = DocumentProcessor()
```

```
docs = processor.load_documents(['file1.txt', 'file2.txt'])
```

```
generator = SyntheticDataGenerator()
```

```
qa_pairs = generator.generate_qa_pairs(docs, num_pairs=50)
```

```
generator.save_benchmark({'qa_pairs': qa_pairs}, 'dataset.json')
```

---

## Key Achievements

- 🎯 **Component Requirement Met** - Synthetic Data Generation fully implemented
- ★ **Outstanding Quality** - 94.4% quality score
- ★ **Excellent Diversity** - 81.9% diversity score
- ✅ **Ethical Compliance** - All protections applied
- 📊 **Ready for Use** - Can generate larger datasets on demand
- 🔄 **Reproducible** - Consistent quality across generations



## KNOWLEDGE BASE (RAG)

### GitHub Links

- [data/sample\\_docs/](#)
- [Knowledge Base README](#)

---

## Overview

The knowledge base contains sample documents used to demonstrate ContextWeaver's RAG capabilities. These documents are specifically chosen to showcase multi-document reasoning, contradiction detection, and temporal analysis.

---

## Knowledge Base Contents

### Sample Documents (3 Files)

#### 1. coffee\_study\_2018.txt

- **Year:** 2018
- **Type:** Research Study

- **Domain:** Medical Research
- **Topic:** Coffee and cardiovascular risk
- **Size:** ~1.2 KB
- **Key Finding:** High coffee consumption (>5 cups/day) increases cardiovascular risk by 23%
- **Limitation:** Did not control for confounding variables (smoking, sugar intake, exercise)
- **Purpose in KB:** Establishes baseline understanding, provides contradiction example

## 2. coffee\_study\_2023.txt

- **Year:** 2023
- **Type:** Research Study
- **Domain:** Medical Research
- **Topic:** Protective effects of moderate coffee
- **Size:** ~1.3 KB
- **Key Finding:** Moderate coffee consumption (2-3 cups/day) shows 15% reduction in cardiovascular risk
- **Strength:** Rigorous controls for confounding variables
- **Purpose in KB:** Demonstrates methodological evolution, resolution of contradiction

## 3. meta\_analysis\_2022.txt

- **Year:** 2022
- **Type:** Meta-Analysis
- **Domain:** Research
- **Topic:** Review of 50 coffee and heart health studies
- **Size:** ~1.1 KB
- **Key Finding:** Explains contradictions in previous research via improved methodology
- **Scope:** Synthesizes findings from 2010-2022
- **Purpose in KB:** Provides context for contradiction resolution, demonstrates synthesis

---

# Knowledge Base Design Rationale

## Why These Specific Documents?

### 1. Temporal Coverage (6 Years)

- 2018 → 2022 → 2023
- Shows knowledge evolution over time
- Tests temporal analysis features
- Demonstrates recency ranking

### 2. Apparent Contradictions

- 2018: "Coffee increases risk"
- 2023: "Coffee reduces risk"
- Perfect for testing contradiction detection
- Shows resolution through better methodology

### 3. Methodological Evolution

- 2018: Uncontrolled study
- 2022: Meta-analysis identifies flaws
- 2023: Controlled study
- Demonstrates how science improves

### 4. Multiple Evidence Levels

- Individual study (2018)
- Meta-analysis (2022) - higher credibility
- Recent study (2023) - high recency
- Tests credibility ranking

### 5. Clear Narrative Arc

- Beginning: Negative findings
- Middle: Critical analysis
- End: Positive findings with explanation
- Perfect for multi-hop reasoning

---

## Knowledge Base Statistics

### Overall Metrics

Total Documents: 3

Total Size: ~3.6 KB

Total Chunks Created: 3 (using hybrid chunking)

Vector Embeddings: 36 total in ChromaDB

Processing Performance:

Loading Time: 0.01s

Embedding Time: 5.70s

Total Ingestion: 9.36s

## Metadata Distribution

### By Domain:

- Research: 100% (3/3 documents)

### By Year:

- 2018: 33% (1 document)
- 2022: 33% (1 document)
- 2023: 33% (1 document)

### By Source Type:

- General Source: 100% (3/3)
- Credibility: Medium (all documents)

### By Quality:

- Average Quality Score: 90.2%
- All documents > 80% quality threshold

### Temporal Coverage:

- Range: 2018-2023 (6 years)
  - Coverage Score: 23.3%
- 

## RAG Component Implementation

### Requirement Checklist

#### ✓ 1. Build a knowledge base for your domain

#### Implementation:

- Hierarchical organization system (**KnowledgeBase** class)
- Documents organized by:
  - Domain (legal, medical, technical, research, financial)
  - Year (temporal indexing)

- Source type (peer-reviewed, preprint, blog, official)
  - Credibility tier (high, medium, low)
- Statistics tracking and coverage scoring
- Metadata enrichment pipeline

**Verified:** 3 documents organized, 23.3% coverage score calculated

---

## ✅ 2. Implement vector storage and retrieval

**Implementation:**

- ChromaDB vector database (persistent storage)
- OpenAI embeddings (text-embedding-3-small, 1536 dimensions)
- Cosine similarity search
- Metadata filtering capabilities
- Collection management

**Verified:** 36 vector embeddings stored, 77.6% similarity achieved in tests

---

## ✅ 3. Design relevant document chunking strategies

**Implementation:** 4 chunking strategies

### Strategy 1: Fixed-Size

- Fixed chunks of 1000 tokens with 200-token overlap
- Uses: Consistent chunk sizes
- Pros: Simple, predictable
- Cons: May split mid-sentence

### Strategy 2: Semantic

- Boundary-aware chunking based on meaning
- Uses: Content-aware splitting
- Pros: Preserves semantic units
- Cons: More complex

### Strategy 3: Sentence-Based

- Respects sentence boundaries
- Uses: Natural language structure
- Pros: No mid-sentence splits
- Cons: Variable chunk sizes

#### Strategy 4: Hybrid ★ (USED IN TESTS)

- Combines semantic boundaries with size constraints
- Uses: Best of both approaches
- Pros: Balanced quality and consistency
- Cons: Slightly more complex
- **Result:** Best performance in testing

**Verified:** All 4 strategies implemented and tested

---

#### ✓ 4. Create effective ranking and filtering mechanisms

##### Implementation: Multi-Factor Ranking

##### 5 Ranking Factors:

1. **Similarity (35%)** - Vector cosine similarity
2. **Credibility (20%)** - Source type quality (peer-reviewed > blog)
3. **Recency (20%)** - Exponential decay with age
4. **Quality (15%)** - Content quality score (structure, completeness)
5. **Alignment (10%)** - Query-document alignment

##### Formula:

$$\text{final\_score} = (\text{similarity} \times 0.35) + (\text{credibility} \times 0.20) +$$
$$(\text{recency} \times 0.20) + (\text{quality} \times 0.15) +$$
$$(\text{alignment} \times 0.10)$$

##### Filtering Mechanisms:

- Year range filtering (year\_min, year\_max)
- Source type filtering (peer-reviewed, preprint, etc.)
- Domain filtering (medical, legal, technical, etc.)
- Quality threshold filtering (min\_quality score)
- Diversity filtering (max per source/year)

**Verified:** Ranking score 0.643 achieved, all filters working

---

## Knowledge Base Organization

## Hierarchical Structure

### Level 1: By Domain

research/

└─ coffee\_study\_2018.txt

└─ coffee\_study\_2023.txt

└─ meta\_analysis\_2022.txt

### Level 2: By Year

2018/ → coffee\_study\_2018.txt

2022/ → meta\_analysis\_2022.txt

2023/ → coffee\_study\_2023.txt

### Level 3: By Source Type

general\_source/

└─ All 3 documents (would be separated in real-world)







### Level 4: By Credibility

medium\_credibility/

└─ All 3 documents

## Automatic Metadata Extraction

### For Each Document:

-  Domain classification (using keyword matching across 5 domains)
  -  Year extraction (from filename or content)
  -  Source type classification (peer-reviewed, preprint, blog, official)
  -  Credibility scoring (based on source type: 0.5-1.0)
  -  Quality scoring (based on length, structure, completeness)
  -  Entity extraction (years, organizations, key phrases)
-

# Knowledge Base Usage

## How Documents Are Loaded

### Process:

1. User provides file paths
2. DocumentProcessor loads files (supports PDF, TXT, DOCX)
3. Text extracted and chunked (using hybrid strategy)
4. Metadata automatically enriched
5. Documents organized hierarchically
6. Chunks stored in ChromaDB with embeddings
7. Statistics calculated and updated

### Result from Test:

- 3 files → 3 chunks in 0.01s
- All metadata extracted successfully
- Knowledge base organized by all 4 hierarchies

## How Documents Are Retrieved

### Process:

1. User submits query
2. Query converted to embedding (1536-dim vector)
3. Similarity search in ChromaDB (cosine distance)
4. Multi-factor ranking applied
5. Metadata filtering (if specified)
6. Graph expansion (if enabled)
7. Top-K documents returned

### Result from Test:

- Query: "Is coffee safe?"
- Retrieved: 4 documents
- Max similarity: 77.6%
- Retrieval time: <1s

---

## Expanding the Knowledge Base

### How to Add New Documents

## Step 1: Add Files

bash

```
cp your_document.pdf data/sample_docs/
```

## Step 2: Ingest

python

```
from src.contextweaver_pipeline import ContextWeaverPipeline
```




```
pipeline = ContextWeaverPipeline()
```

```
pipeline.ingest_documents(['data/sample_docs/your_document.pdf'])
```

## Step 3: Verify

- Check knowledge base statistics
- Test query on new content
- Verify embeddings created

## Supported File Formats

-  PDF (.pdf) - Via pypdf library
-  Text (.txt) - Direct read
-  Word (.docx) - Via python-docx library

---

# Knowledge Base Quality

## Current Quality Metrics

**Coverage Score:** 23.3%

- Calculation:  $(\text{domain\_diversity} + \text{temporal\_diversity} + \text{source\_diversity}) / 3$
- 23.3% indicates: Small but focused knowledge base
- For full coverage: Need 5+ domains, 10+ years, 5+ source types

**Document Quality:** 90.2% average

- All documents above 80% quality threshold





- High structure scores
- Complete content (no truncation)

**Credibility:** Medium tier

- All documents classified as "general\_source"
  - Would improve with peer-reviewed sources
  - Suitable for demonstration purposes
- 

## RAG Component Achievements

 **All 4 RAG Requirements Met:**

1.  Knowledge base built with hierarchical organization
2.  Vector storage implemented (ChromaDB + OpenAI embeddings)
3.  Chunking strategies designed (4 strategies, hybrid performs best)
4.  Ranking mechanisms created (5-factor multi-factor ranking)

 **Beyond Requirements:**

- Knowledge graph integration (NetworkX)
- Graph-based retrieval (not just similarity)
- PageRank document importance
- Advanced metadata enrichment
- Automatic domain classification

 **Test Validation:**

- 100% of RAG tests passed
- 77.6% similarity accuracy
- 0.643 ranking score
- 0.01s processing speed