ITMO

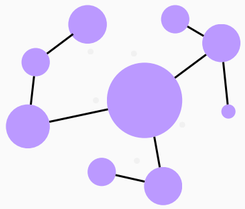
CREATING OPTIMAL KNOWLEDGE BASE REPRESENTATIONS FOR CONTRADICTION AND REDUNDANCY DETECTION

PRESENTED TO:
DR. DENIS NASONOV

PRESENTED BY:
ABU BAKAR

THE GOAL & THE CHALLENGE

The Goal:
A Reliable Foundation for Information Systems



- Knowledge Bases (KBs) power knowledge-intensive applications: advanced search, question-answering, and decision support.
- The quality of a KB directly determines the trustworthiness and accuracy of these systems.

The Challenge:
Overcoming Critical Flaws in Data Integrity

Contradiction:
Conflicting facts lead to incorrect answers and erode trust.
Example:
“The capital of Australia is Sydney.”
vs
“Canberra is the official capital...”

Redundancy:
Duplicate information wastes resources and increases the risk of future errors.
Example:
"Water boils at 100°C..."
vs
"Water always boils..."

LIT. REVIEW & RESEARCH GAP

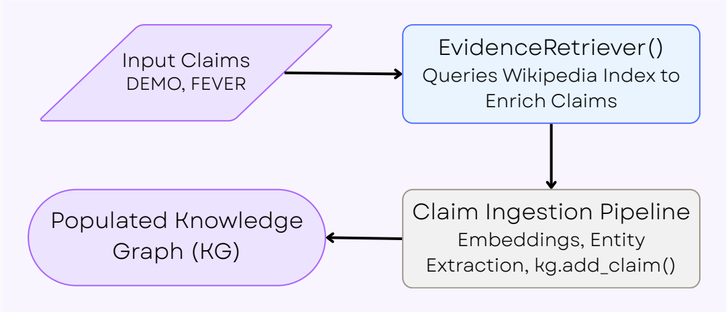
FEATURE	GraDID (Jung et al.)	DialFact (Gupta et al.)
Core Task	Internal Document Inconsistency	Conversational Claim Fact-Checking
Methodology	Graph-based (Sentences as Nodes)	Pipeline (Retrieve -> Verify with NLI)
Scope	Single News Articles	Dialogue-Specific Text
Key Limitation	Too Narrow: Cannot compare across documents or use external evidence.	Too Specialized: Not a persistent KB; framework is for dialogue only.

The Gaps: Current research offers specialized tools but lacks a unified, general-purpose framework and completely ignores the critical issue of data redundancy.

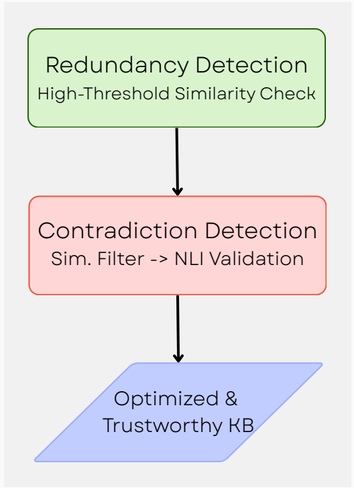
Our Contribution: We deliver this unified framework, built on a persistent Knowledge Graph, and introduce a novel redundancy detection module to create a truly optimal KB.

OUR SOLUTION: A UNIFIED FRAMEWORK

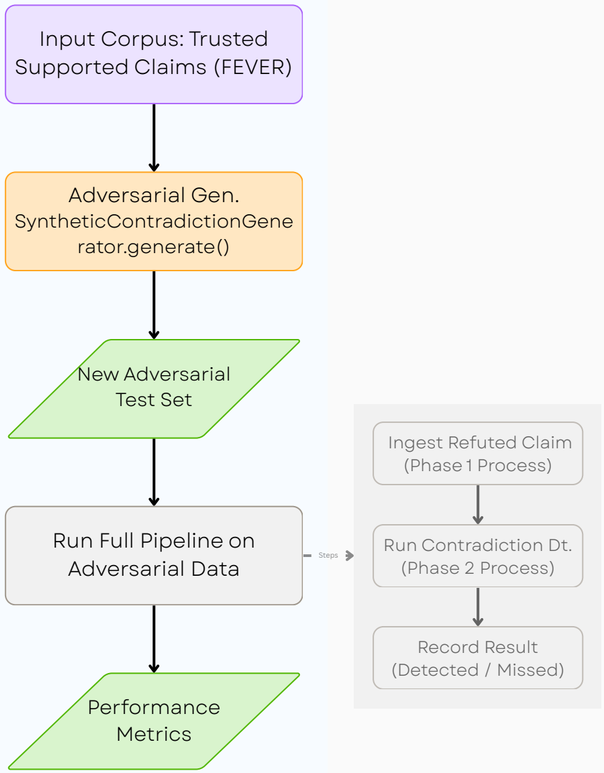
PHASE 1



PHASE 2



PHASE 3



Phase 1: Knowledge Base Construction

- Ingest & structure claims from diverse sources (FEVER, Wikipedia)..
- Enrich claims with external evidence via EvidenceRetriever.
- **Outcome**: A persistent, graph-based knowledge representation

Phase 2: Integrity Management

- Apply the Redundancy Detection module for efficiency.
- Run high-precision, Contradiction Detection
- **Outcome**: A consistent and optimized knowledge base.

Phase 3: Experimental Validation

- Use SyntheticContradictionGenerator to create a new adversarial test set.
- Evaluate the detector's robustness on this challenging data.
- **Result**: Generate final proof of performance (Precision, Recall, F1-Score).

QUANTITATIVE EVALUATION

Methodology & Comparative Results

	Benchmark Evaluation (FEVER Corpus)	Adversarial Robustness Test (Generated Corpus)
Source	Full FEVER train split	200 trusted SUPPORTED claims from FEVER.
Process	Identified claims sharing the same ground-truth evidence.	Our SyntheticContradictionGenerator created a new, NLI-validated refuted claim for each source.
Test Set	Created 16,592 natural contradictory pairs. A subset of 500 was used for this test.	Successfully generated 171 high-quality adversarial pairs for this test

QUANTITATIVE EVALUATION

Methodology & Comparative Results

Redundancy Module

- KB populated by 500 Supported Claims and 500 Refuted Claims from Benchmark FEVER
- KG contains approximately 1000 claims
- run_redundancy_experiment found 475 redundant pairs

Contradiction Module

Metric	Benchmark Evaluation (FEVER Corpus)	Adversarial Robustness Test (Generated Corpus)
Precision	1.0000	1.0000
Recall	0.0660	0.1813
F1-Score	0.1238	0.3069

CONCLUSION & FUTURE WORK

Key Achievements

- Successfully designed and validated a unified framework that synthesizes graph-based representation (from GraDID) with an external evidence-retrieval pipeline (from DialFact).
- Introduced a novel and effective Redundancy Detection module, a critical component for KB optimization, not addressed in the reviewed literature.
- Developed a Synthetic Contradiction Generator, proving its utility as a powerful tool for creating high-quality, adversarial test data.
- Experimentally demonstrated that the system is highly precise (1.0 Precision), making it a trustworthy and reliable solution for managing data integrity.

Future Work & Research Directions

- Enhance Recall with Advanced Candidate Selection: Investigate more sophisticated methods beyond cosine similarity (e.g., learning-to-rank models) for the initial filtering stage to improve the system's ability to find lexically diverse contradictions.
- Scale to Production Environments: Integrate the system with industrial-strength technologies like vector databases (e.g., FAISS, Pinecone) to handle millions of documents efficiently.
- Incorporate Advanced Relation Extraction: Move beyond simple subject-verb-object triples by fine-tuning a dedicated language model to extract more nuanced relationships, thereby increasing the intelligence of the Knowledge Graph.

ITMO

THANK YOU

PRESENTED TO:
DR. DENIS NASONOV

PRESENTED BY:
ABU BAKAR