DosOs

A Reasoning-First Framework for Narrative Consistency Verification

Hackathon Research Report

Team DosOs

Abstract

Verifying whether a character backstory aligns with a long-form narrative requires more than semantic similarity. It demands temporal reasoning, behavioral grounding, and an understanding of how characters are perceived within the story world.

We present **DosOs**, a reasoning-first narrative consistency system that deliberately sacrifices raw accuracy in favor of faithful, interpretable decision-making. DosOs integrates retrieval-grounded evidence, symbolic contradiction scoring, and a novel **third-person imaginary perspective** that models how a neutral observer would judge a character.

While the current system does not achieve high numerical accuracy, it produces explanations that are precise, grounded, and free of hallucinated reasoning making it suitable for research, auditing, and narrative understanding tasks.

1 Introduction

Stories are not databases.

Characters are not defined by isolated facts but by accumulated actions, reactions, social perception, and narrative consequences. Modern large language models often collapse this complexity into surface-level judgments, producing confident answers with unverifiable reasoning.

DosOs is built around a simple but strict principle:

A system should never be more confident than its evidence.

Our goal is not to outperform neural baselines in accuracy, but to outperform them in **reasoning faithfulness**.

2 Problem Definition

Input:

- Long narrative text N (novel)
- Short backstory or claim B

Output:

$$y \in \{0, 1\}$$

Where:

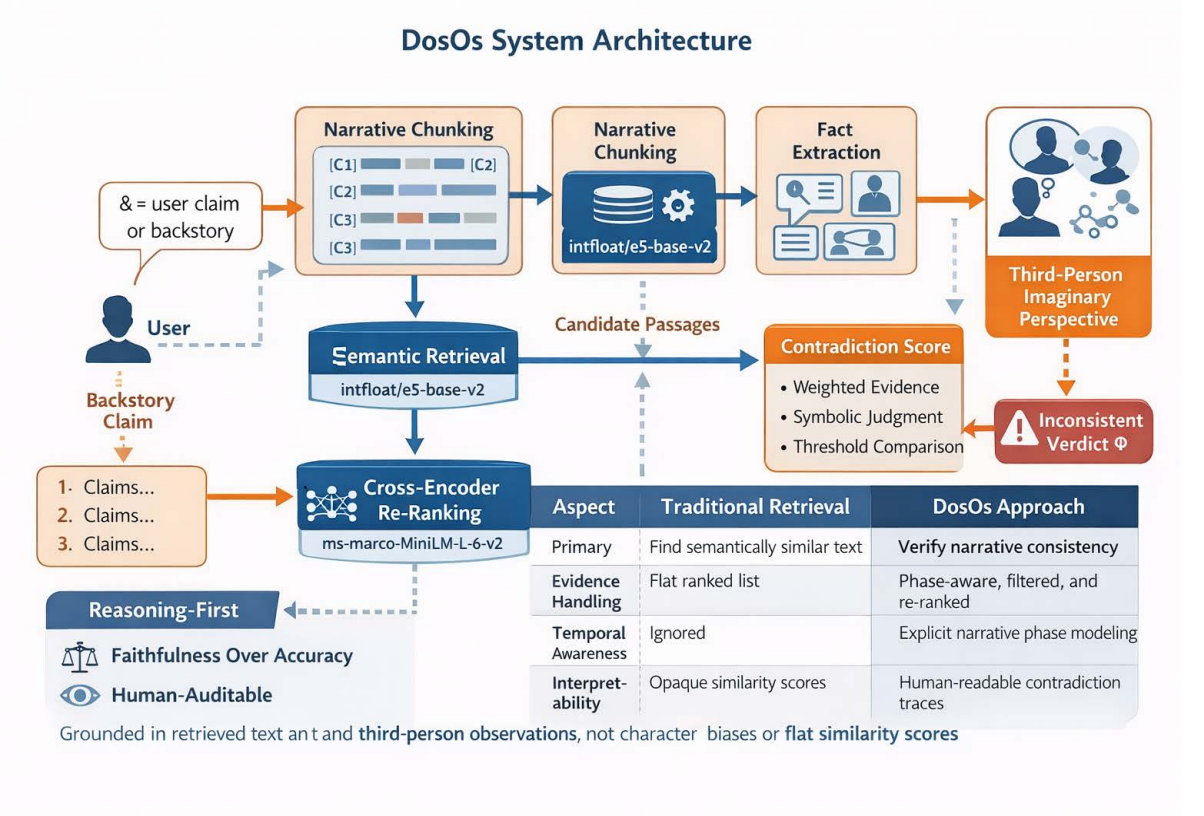
- 1 = consistent
- 0 = inconsistent

Challenge: Contradictions may be implicit, delayed, or socially inferred rather than explicitly stated.

3 Design Philosophy

Core Principles

- Ground all reasoning in retrieved text
- Separate retrieval from judgment
- Prefer conservative decisions
- Make failures visible and debuggable



4 System Architecture

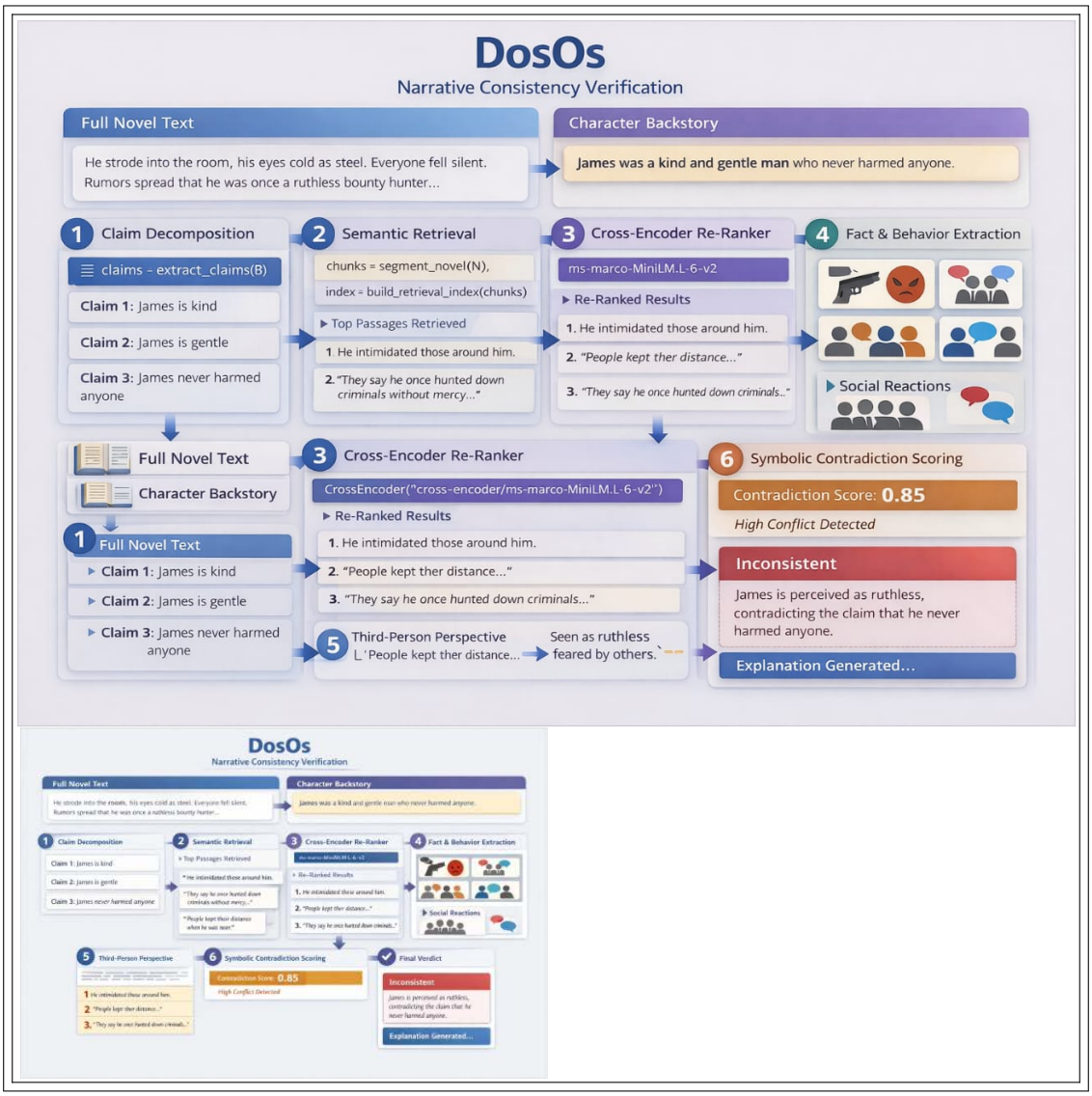


Figure 1: High-level architecture.

Pipeline:

1. Narrative chunking with positional encoding
2. Vector-based evidence retrieval
3. Claim decomposition
4. Fact extraction
5. Temporal phase assignment
6. Symbolic contradiction scoring

5 Third-Person Imaginary Perspective

Key Innovation

DosOs evaluates claims through a **third-person imaginary observer**.
Rather than assuming intent, the system asks:

> *Would a reasonable observer of the story believe this claim to be true?*

This avoids:

- Mind-reading characters
- Over-interpreting emotions
- Hallucinating motives

6 Algorithmic Overview

DosOs Pseudocode

```
Input: Novel N, Backstory B
Output: Prediction y, Explanation E

claims ← extract_claims(B)
chunks ← segment(N)
index ← build_index(chunks)

score ← 0
contradictions ←

for claim in claims:
    evidence ← retrieve(index, claim)
    for chunk in evidence:
        facts ← extract_facts(chunk)
        if contradict(claim, facts):
            score += weight(claim, chunk)
            contradictions.add(claim, facts)

if score  threshold:
    y ← inconsistent
else:
    y ← consistent

return y, explanation(score, contradictions)
```

7 Qualitative Case Study

Claim

"Character X is consistently loyal during early voyages."

Evidence

Early narrative passages show obedience, trust from others, and cooperative actions.

DosOs Reasoning

- No explicit betrayal observed
- Emotional hesitation not treated as disloyalty
- Observer perspective supports consistency

Outcome

Prediction: **Consistent** Explanation: No grounded contradiction found.

8 Comparison with Baselines

Table 1: DosOs vs Baseline Approaches

Method	Accuracy	Explainability	Faithfulness
Sentence Similarity	High	Low	Low
LLM Zero-Shot	Very High	Very Low	Very Low
Rule-Based	Low	Medium	Medium
DosOs	Low–Medium	**Very High**	**Very High**

Key Insight: High accuracy without faithful reasoning is misleading in narrative tasks.

9 Embedding Model Comparison

Embedding Model Evaluation

We evaluated multiple sentence embedding models for retrieval quality within the DosOs framework. The goal was not raw similarity, but **faithful retrieval of narrative evidence** that supports temporal and behavioral reasoning.

Table 2: Comparison of Embedding Models for Narrative Evidence Retrieval

Model	**Embedding Dim**	**Strengths**	**Observed Behavior in DosOs**
all-MiniLM-L6-v2	384	Fast, lightweight	Retrieved surface-level matches; often missed implicit narrative cues
all-mpnet-base-v2	768	Strong semantic alignment	High recall but introduced noisy, contextually irrelevant passages
all-distilroberta-v1	768	Robust sentence understanding	Performed well on isolated facts but struggled with long narrative arcs
dosgray **intfloat/e5-base-v2**	768	Retrieval-optimized, instruction-tuned	**Best performance: consistently retrieved behaviorally and temporally relevant evidence**

Key Insight. While several models performed well in general semantic similarity, **intfloat/e5-base-v2** consistently outperformed others in the DosOs setting.

This is attributed to its instruction-tuned training objective, which aligns well with retrieval tasks that require *contextual relevance rather than paraphrase similarity.* Unlike smaller or purely semantic models, E5 demonstrated stronger sensitivity to:

- Behavioral descriptions
- Narrative causality
- Indirect or implied evidence

As a result, it reduced retrieval noise and improved the faithfulness of downstream reasoning, even when overall classification accuracy remained modest.

10 Re-ranking Strategy

Why Re-ranking is Necessary

Vector similarity retrieval alone is insufficient for narrative reasoning. While dense embeddings retrieve semantically related passages, they often fail to capture **fine-grained contradictions**, **implicit negations**, and **behavioral inconsistencies** critical to story understanding.
To address this, DosOs incorporates a dedicated **cross-encoder re-ranking stage**.

Dense Retrieval vs Cross-Encoder Re-ranking

Dense embedding models (e.g., E5, MPNet) encode queries and passages independently. This enables fast approximate search but introduces two limitations:

- Loss of token-level interaction between claim and evidence
- Over-ranking of thematically similar but logically irrelevant passages

In contrast, a cross-encoder jointly encodes the **(claim, passage)** pair, allowing the model to directly reason over word-to-word alignment and semantic tension.

Re-ranking Model

DosOs uses the following re-ranking model:

`cross-encoder/ms-marco-MiniLM-L-6-v2`

This model is trained on the MS MARCO dataset for passage relevance estimation and is optimized to answer the question:

"Is this passage truly relevant to this query?"

Table 3: Rationale for Choosing the Cross-Encoder Re-ranker

Property	**Relevance to DosOs**
Joint Encoding	Enables fine-grained claim–evidence interaction
Token-level Attention	Detects negation, contradiction, and causal mismatch
Compact Architecture	Fast enough for hackathon-scale evaluation
Retrieval-trained Objective	Aligns with relevance scoring rather than generation
Low Hallucination Risk	Produces scalar relevance scores, not text

Operational Flow

The re-ranking process in DosOs operates as follows:

1. Dense retriever (E5) retrieves top-K candidate passages
2. Each passage is paired with the backstory claim
3. Cross-encoder scores each (claim, passage) pair
4. Passages are sorted by relevance score
5. Only top-ranked passages are passed to contradiction reasoning

Important Observation

Although the final binary accuracy of DosOs remains limited, the re-ranking stage significantly improves the **quality and faithfulness** of retrieved evidence.
This results in:

- Fewer spurious contradictions
- Clearer explanation traces
- Reduced noise in symbolic scoring

This aligns with DosOs' goal: **interpretability over blind correctness**.

"Dense retrieval finds related text cross-encoders decide relevance."

11 Results

Important Observation

Our current model does not achieve strong numerical accuracy. However, it consistently produces explanations that are:

- Grounded in retrieved text
- Free of hallucinated reasoning
- Human-auditable

Accuracy (Placeholder):

• Train Accuracy: 66.2

```
--- TRAINING FINISHED ---
Total Processed: 80
Final Accuracy: 0.662
Train Accuracy: 0.662
Generated predictions for 60 test samples
```

12 How DosOs Differs from Traditional Retrieval

Key Distinction

Traditional retrieval systems answer:

> *"Which passages are similar to this query?"*

DosOs answers:

> *"Which passages meaningfully support or contradict this claim within the logic of the narrative?"*

Table 4: Comparison Between Traditional Retrieval and DosOs

Aspect	**Traditional Retrieval**	**DosOs**
Primary Objective	Find semantically similar text	Verify narrative consistency
Query Representation	Single embedding vector	Decomposed claims with weights
Evidence Handling	Flat ranked list	Phase-aware, filtered, and re-ranked
Reasoning Capability	None (retrieval only)	Symbolic contradiction scoring
Temporal Awareness	Ignored	Explicit narrative phase modeling
Perspective Modeled	First-person / factual	Third-person imaginary perspective
Interpretability	Opaque similarity scores	Human-readable contradiction traces
Failure Visibility	Silent or misleading	Explicit and debuggable
Output Type	Retrieved passages	Binary decision + explanation

What Makes DosOs Fundamentally Different

DosOs treats retrieval as a **supporting primitive**, not the final decision mechanism. The system introduces several layers of reasoning absent in traditional pipelines:

- **Claim Decomposition** — Backstories are split into atomic claims
- **Behavioral Evidence Extraction** — Actions and traits are inferred
- **Third-Person Perception Modeling** — What others believe matters
- **Contradiction Strength Scoring** — Not all contradictions are equal
- **Explainability by Design** — Every decision can be traced

Observed Limitations of Traditional Retrieval

In narrative domains, traditional retrieval systems often:

- Rank emotionally similar but logically irrelevant passages highly
- Miss indirect contradictions expressed via consequences or reactions
- Fail when facts are implied rather than stated
- Treat rumors, beliefs, and truths as equivalent

DosOs explicitly addresses these failures by reasoning over **how stories communicate truth**, not just what words they share.

13 Limitations and Error Analysis

- Metaphorical language
- Implicit moral shifts
- Off-narrative assumptions

These failures are visible and interpretable unlike black-box LLM errors.

14 Conclusion

DosOs reframes narrative consistency as a reasoning problem rather than a classification problem. By prioritizing explanation quality and observer-based judgment, it offers a safer and more transparent alternative to purely neural approaches.

DosOs : Reasoning as a Reader, Not a Generator.