0.7907 f1 Score on Train with Cross Valid

Global Narrative Consistency Checking via Evidence-Grounded NLI Reasoning

Kharagpur Data Science Hackathon 2026

Date: 11, January 2025

Team Name
hariswarsamasi

Problem Overview

Large Language Models (LLMs) perform well on localized text understanding tasks such as summarization and question answering. However, they often fail to maintain global consistency over long narratives, where meaning emerges cumulatively through events, character development, and causal constraints.

The task in this challenge is framed as a binary decision problem. Given:

- the complete text of a long-form narrative (a novel),
- and a newly proposed hypothetical backstory for a central character,

the system must determine whether the backstory is globally consistent with the narrative as a whole.
This task does not require text generation or thematic interpretation.
Instead, it requires:

- aggregation of evidence distributed across long contexts,
- tracking evolving narrative constraints,
- and causal reasoning over time.

All experiments and evaluations were executed in a Kaggle Notebook environment using publicly available models and datasets to ensure full reproducibility. Code link in Last Page

⚙ System Architecture

Our system follows a retrieval-augmented reasoning pipeline, designed to operate robustly over long narratives without truncation or summarization.

1.Data Ingestion

- Full novels are ingested in raw .txt format without truncation.
- Each book is processed independently to preserve narrative boundaries.
- Metadata such as book_name, character, and backstory content are retained.

2.Long-Context Handling

To manage 100k+ word narratives, we use sliding-window chunking:

- Chunk size: 800 words
- Overlap: 200 words
- This preserves local coherence while ensuring global coverage.

3.Evidence Retrieval

For each backstory:

1. Dense embeddings are generated using a SentenceTransformer model.
2. The backstory text is embedded and compared against all chunks from the corresponding novel.
3. Top-K most similar chunks (K = 2) are retrieved as candidate evidence.

This design ensures:

- scalability to long texts,
- minimal noise from irrelevant sections,
- and retrieval grounded in semantic similarity rather than keyword overlap.

Reasoning & Consistency Scoring

1.Natural Language Inference

- Each (backstory, evidence) pair is evaluated using a pretrained MNLI model:
- Model: roberta-large-mnli
- Output labels: ENTAILMENT, NEUTRAL, CONTRADICTION
- This allows the system to assess logical compatibility, not surface plausibility.

2.Evidence Aggregation Strategy

For a given backstory, scores from multiple evidence chunks are aggregated:

- ENTAILMENT increases confidence in consistency.
- CONTRADICTION is weighted more heavily, reflecting irreversible narrative violations.
- NEUTRAL contributes a small negative penalty to discourage weak matches.

A soft-veto mechanism is applied when strong contradictions are detected, ensuring that even a single decisive inconsistency can dominate the final judgment.

3.Final Decision Rule

The aggregated score is compared against an F1-optimized decision threshold learned from training data.

- Score ≥ threshold → Consistent (1)
- Score < threshold → Contradict (0)

This framing converts long-context reasoning into a structured classification problem, as required by the task.

Experimental Setup

1.Cross-Validation

- 5-fold stratified cross-validation
- Stratification preserves class balance across folds.
- All hyperparameters and thresholds are selected using only training folds.

2.Models Evaluated

We evaluated multiple NLI backbones:

Model	Mean F1	Mean Accuracy
roberta-large-mnli	0.7907	0.6625
microsoft/deberta-v3-large	0.7783	0.6375
facebook/bart-large-mnli	0.7441	0.600

roberta-large-mnli consistently achieved the best performance and stability.

3.Ablation Studies

We tested several extensions:

- contradiction consensus (requiring multiple contradictions),
- character-specific weighting,
- symbolic irreversibility rules (e.g., “never”, “first time”).

None of these provided consistent improvements over the base system, indicating that the learned NLI-based reasoning already captured these constraints. We therefore retained the simpler and more robust model.

Final Decision Rule

Cross-Validation Performance

- Mean Accuracy: 0.6625
- Mean F1 Score: 0.7907

This demonstrates strong robustness and generalization despite limited labeled data.

Limitations

While effective, the system has known limitations:

1. **Implicit Causality:**
   Some contradictions rely on unstated world knowledge or subtle temporal assumptions that pretrained NLI models may miss.
2. **Small Dataset:**
   The limited number of labeled examples restricts fine-tuning and favors pretrained reasoning models.
3. **Narrative Diversity:**
   The dataset contains only a small number of novels; broader genres may require recalibration.

Conclusion

We present a retrieval-augmented, evidence-grounded reasoning system for global narrative consistency checking. By avoiding text generation and focusing on structured causal reasoning, the system aligns closely with the challenge's evaluation goals.

Our approach demonstrates that careful aggregation of distributed evidence, combined with pretrained NLI models, can outperform locally coherent but globally inconsistent reasoning strategies.

The system is fully reproducible, robust, and interpretable, making it suitable for real-world long-context reasoning tasks.

Links

[Hariswar8018/IIT Kharagpur Submission Repo]

[Training Kaggle Notebook]

Author

Team Name : hariswarsamasi

Member 1 : Ayusman Samasi (Solo Team)

Email : hariswarsamasi@gmail.com