Kharagpur Data Science Hackathon 2026

Track A: Systems Reasoning with NLP and Generative AI

Team: The Auralion

---

1. Introduction

Large language models often struggle to maintain global consistency when reasoning over long narratives. While they may perform well on short passages, they frequently fail to account for constraints formed earlier in a story, leading to logically inconsistent conclusions.
This hackathon focuses on evaluating such failures by testing whether a hypothetical backstory of a character can logically and causally fit within a complete long-form narrative.

---

2. Problem Statement

The task is framed as a **binary classification problem**.
Given:

- A narrative derived from a long-form novel
- A hypothetical backstory for a central character

The system must predict:

- **1 (Consistent):** The backstory is logically compatible with the narrative
- **0 (Contradict):** The backstory violates narrative constraints or causal logic

The emphasis is on reasoning over extended context rather than surface-level textual similarity.

3. Dataset Description

The dataset consists of narrative segments derived from long novels (100k+ words) and associated hypothetical backstories.
Each example contains:

- Narrative content related to a character
- A backstory description (caption)
- A ground-truth consistency label for training data

The backstories are intentionally designed to be plausible while potentially conflicting with the narrative, making the task challenging.

---

4. Track Selection

We selected **Track A: Systems Reasoning with NLP and Generative AI**.
This track aligns with our focus on:

- Robust classification
- Evidence-based reasoning
- Well-structured NLP pipelines
  rather than developing new model architectures.

---

5. System Overview

Our solution follows a **feature-based classification pipeline** designed to detect consistency signals between narrative context and backstory.

High-level flow:

Narrative Content → Feature Extraction → Classification → Prediction Output

The system avoids text generation and instead relies on structured linguistic and semantic signals

6. Long-Context Handling

Directly processing entire novels is computationally impractical.
Instead, our system operates on **narrative segments derived from long-form texts**, which preserve relevant contextual information.

This approach allows the model to:

- Capture global consistency cues indirectly
- Avoid memory limitations
- Maintain efficiency while reasoning over long narratives

---

7. Reasoning and Feature Extraction

For each narrative–backstory pair, the following features are extracted:

- Semantic similarity using spaCy embeddings
- Vector distance between narrative and backstory representations
- Named entity overlap
- Negation mismatch between texts
- Absolute-claim mismatch (e.g., “always”, “never”)
- Presence or absence of backstory information

These features collectively capture semantic alignment and logical tension between the narrative and the proposed backstory.

---

8. Classification Model

A **Logistic Regression** classifier is used for prediction.
Key characteristics:

- Lightweight and interpretable
- Balanced class weighting to handle label imbalance
- Trained on extracted linguistic and semantic features

The model outputs a binary decision indicating consistency or contradiction.

9. Use of Pathway Framework

Pathway is integrated as part of the **data ingestion and orchestration layer**.
It is used to structure narrative data and manage processing flow where available.
In environments where Pathway is not supported, the system safely falls back to local execution while preserving identical prediction logic.
This satisfies the Track A requirement of meaningful Pathway usage.

---

10. Output Format

The final output is a CSV file named results.csv containing:

- Story ID
- Binary prediction (1 = consistent, 0 = contradict)

This format aligns exactly with the hackathon submission guidelines.

---

11. Reproducibility

The system is designed to run end-to-end in a clean environment.
All dependencies are explicitly defined, and no manual steps are required once the input data is provided.
A minimal frontend is included only for demonstration and does not affect evaluation.

---

12. Limitations

- The system focuses on classification robustness rather than generating detailed explanations.
- Highly implicit or ambiguous causal chains may be challenging.
- Full symbolic reasoning over entire novels is outside the current scope.

13. Team Contributions

Team Name: The Auralion

- **Bhavya** – Exploratory Data Analysis, feature engineering, and model development
- **Rutvi** – Frontend design and visualization
- **Kena** – Repository management, documentation, report writing, and presentation
- **Digvijay** – Backend support, API integration, and project coordination

Each member contributed to building a cohesive and reproducible solution aligned with Track A objectives.

---

14. Conclusion

This project demonstrates a structured NLP-based approach to evaluating narrative consistency over long contexts. By focusing on feature-driven reasoning, robustness, and reproducibility, our system aligns closely with the goals of Track A and provides a solid foundation for further exploration in long-form narrative reasoning.