

KHARAGPUR DATA SCIENCE **HACKATHON, 2026**



Project Report
on
Track A: Systems Reasoning with NLP and Generative AI

**Title : Evaluating Narrative Backstory Consistency
through Long-Context Reasoning**

(11th January 2026)

Team Name : Tech Monarchs

Aryan Singh

Anshuman Singh

Yash Singh

ABSTRACT

Large Language Models (LLMs) have made great strides in understanding local text tasks like summarization, question answering, and short-context reasoning. However, these models often struggle with long-form narratives, where meaning comes from the buildup of events, decisions, and limits over time. In these narratives, earlier events can limit what happens later, characters change, and causal connections develop or get ruled out. This issue is especially clear in tasks that need to track constraints over time. A backstory that seems reasonable in a small part of the narrative can still be inconsistent with the overall story when looked at across its entire timeline. Current systems often depend on surface-level similarities or local coherence, leading to choices that sound plausible on their own but clash with the full narrative.

The Kharagpur Data Science Hackathon 2026 aims to tackle this problem. Instead of asking participants to create text or interpret themes, the challenge is presented as a decision-making problem: given a complete novel and a fictional backstory for a character, determine if the backstory can logically and causally lead to the narrative that unfolds. This approach highlights the importance of reasoning quality, gathering evidence, and maintaining overall consistency rather than just language fluency.

Our work addresses this challenge by developing a reasoning-first system that explicitly models narrative constraints, collects evidence, and ensures consistency over time. The aim is not to craft elegant explanations but to make careful, evidence-based decisions that remain consistent across longer contexts.

PROBLEM STATEMENT

Each input example in the dataset consists of two primary components. The first is a complete long-form narrative, typically exceeding 100,000 words, provided without truncation or summarization. The second is a hypothetical backstory for a central character, describing early-life events, beliefs, fears, ambitions, and assumptions about the world. This backstory is deliberately designed to be plausible, yet potentially inconsistent with the narrative.

The system must output a binary decision:

- 1** if the backstory is consistent with the narrative, and
- 0** if the backstory contradicts the narrative.

The central difficulty of this task lies in several intertwined challenges. First, the extreme length of the narratives makes naïve processing infeasible. Second, relevant evidence may be distributed across distant parts of the story rather than appearing in a single passage. Third, causal compatibility must be distinguished from superficial plausibility. A backstory may “sound right” while still violating narrative constraints established earlier or later in the text.

Importantly, this is not a language generation task. Producing fluent explanations is insufficient if the underlying reasoning is flawed. Instead, the task demands careful aggregation of evidence, explicit handling of contradictions, and global coherence across narrative time. These requirements motivate a system design that prioritizes structured reasoning over end-to-end generation.

OVERALL SYSTEM APPROACH

To address the challenges outlined above, we design a modular, interpretable reasoning pipeline. Rather than relying on a single monolithic model, the system decomposes the task into well-defined stages, each corresponding to a specific reasoning function.

At a high level, the pipeline operates as follows. The full novel is first ingested and transformed into a long-context semantic memory through chunking and embedding. The hypothetical backstory is then decomposed into a small set of atomic factual claims. For each claim, relevant evidence is retrieved from across the novel. These evidence fragments are evaluated individually, aggregated to avoid cherry-picking, and combined into a final consistency judgment.

This modular design has several advantages. It allows each stage to be inspected and debugged independently, supports evidence-grounded decisions, and makes the reasoning process transparent to evaluators. Most importantly, it mirrors how a human analyst would approach the problem: by breaking down complex narratives into testable claims and checking them against multiple pieces of evidence.

By framing the task as structured reasoning rather than text generation, the system remains robust to noise and avoids overfitting to stylistic cues.

HANDLING LONG CONTENT

Handling extremely long narratives is a core requirement of this task. Our system ingests the complete novel text without truncation, ensuring that all narrative constraints remain accessible throughout the reasoning process. This design choice is essential for maintaining global coherence and enabling consistency checks across the full story arc.

To make processing tractable, the novel is divided into overlapping word-based chunks. Each chunk contains approximately 1000 words, with an overlap of around 200 words between consecutive chunks. Overlap is crucial because many narrative events span paragraph or chapter boundaries. Without overlap, important causal links could be lost, leading to fragmented or misleading evidence retrieval.

Each chunk is embedded into a dense semantic vector representation and stored in a vector memory structure. This long-context memory enables meaning-based retrieval across the entire novel, rather than restricting reasoning to a narrow local window. As a result, the system can retrieve evidence from early, middle, or late stages of the narrative as needed, supporting global reasoning and constraint tracking.

BACKSTORY DECOMPOSITION INTO CLAIMS

Hypothetical backstories often bundle multiple assumptions into a single paragraph. For example, a backstory may simultaneously assert facts about upbringing, long-term beliefs, motivations, and personality traits. Treating such a backstory as a single unit risks masking localized contradictions.

To address this, we decompose each backstory into a small number of atomic factual claims, typically between three and five. Each claim represents a single assertion that can be independently evaluated against the narrative. Examples include statements about childhood experiences, persistent fears, or enduring attitudes toward authority or relationships.

This claim-level decomposition transforms the problem from vague plausibility assessment into a structured reasoning task. Each claim becomes a hypothesis that must be supported, contradicted, or left unconstrained by the narrative evidence. This approach significantly improves interpretability and enables fine-grained error analysis.

EVIDENCE RETRIEVAL AND NOISE CONTROL

For each extracted claim, the system retrieves multiple semantically relevant chunks from the long-context memory. Retrieving multiple evidence fragments ensures that decisions are not based on isolated passages. This enables **multi-evidence aggregation to avoid cherry-picking**, a critical requirement for robust narrative reasoning.

Each retrieved chunk is evaluated independently and categorized as SUPPORT, CONTRADICT, or UNCLEAR with respect to the claim. These labels force explicit judgments rather than relying on raw similarity scores. By aggregating verdicts across multiple chunks, the system reduces sensitivity to noisy retrieval results and avoids over-reliance on any single piece of evidence.

If evidence fragments provide mixed signals, the system conservatively treats the claim as uncertain. This conservative bias helps prevent false contradictions or false support arising from ambiguous passages.

WEIGHTED REASONING & TEMPORAL CONSISTENCY

Narrative reasoning is inherently asymmetric. A single strong contradiction often outweighs several weakly supportive passages. To reflect this, we apply **asymmetric penalties for contradictions** in the claim scoring process. Support contributes positively, uncertainty contributes neutrally, and contradictions incur a larger negative penalty.

Beyond static aggregation, the system also enforces temporal consistency. Evidence chunks are associated with their relative position in the narrative, allowing the system to reason about how character traits and beliefs evolve over time. This enables detection of temporal inconsistencies, such as backstories that assert permanent traits contradicted by early or late narrative developments.

This temporal analysis is a key mechanism for **constraint tracking over narrative time**, ensuring that the final judgment reflects the full evolution of the story rather than isolated snapshots.

SELECTIVE GENERATIVE REASONING & FINAL DECISION

While generative models are powerful, we deliberately avoid end-to-end text generation. Instead, we adopt **selective generative reasoning, not end-to-end generation**, where generative components are optionally used only for localized decisions such as assessing claim–evidence compatibility. The overall pipeline remains deterministic and interpretable.

After aggregating scores across all claims, the system applies a simple final decision rule. If the total score is negative, the backstory is labeled inconsistent; otherwise, it is labeled consistent. This rule is transparent, easy to justify, and robust to noise.

Predictions are written to a results.csv file with strict guarantees: correct column names, no NaN values, no extra columns, and one row per test example. This ensures compatibility with automated evaluation and avoids disqualification due to formatting errors.

LIMITATIONS & CONCLUSION

Despite its strengths, the system has important limitations. The system may miss subtle psychological contradictions not explicitly stated. Implicit emotional shifts or unstated motivations may not be captured by claim-based reasoning. Additionally, heuristic claim extraction may overlook nuanced assumptions, and temporal reasoning relies on chunk position rather than explicit event graphs.

These limitations reflect the inherent difficulty of deep narrative reasoning and point toward future work involving richer event representations and more sophisticated causal modeling.

In conclusion, we present a reasoning-driven system for evaluating backstory consistency over long narratives. By combining long-context memory, structured claim decomposition, multi-evidence aggregation, weighted contradiction handling, and temporal consistency checks, our approach directly addresses the core challenges of Track A. The system emphasizes global coherence and evidence-grounded reasoning over surface-level plausibility, aligning closely with the goals of the hackathon.