Motivation

Large language models perform well on local text understanding tasks such as summarization and question answering. However, they struggle with global consistency over long narratives, where meaning emerges from how events, states, and constraints accumulate over time.

In long-form text, earlier events restrict what can plausibly happen later. Characters change, commitments are made, and causal pathways are either reinforced or ruled out. Correct reasoning in this setting requires **tracking how these constraints evolve** and determining whether a given future is compatible with a proposed past.

Current models often fail at this type of reasoning. They rely on surface-level plausibility, producing explanations that are locally coherent but globally inconsistent. As a result, they confuse correlation with causation and narrative similarity with logical compatibility.

This challenge is designed to evaluate that failure mode. Participants are not asked to generate text or interpret themes. Instead, the core task is a decision problem: given evidence distributed across a long narrative, **determine whether a hypothesized past can causally and logically produce an observed future**.

Although the task is framed in narrative terms, it ultimately reduces to a **structured classification problem over long contexts** - requiring careful evidence aggregation, constraint tracking, and causal reasoning rather than language generation.

About Pathway

Pathway is building the world's first frontier model for enterprise.
Its breakthrough architecture - **The Dragon Hatchling** outperforms transformers and provides enterprises with full visibility into how the model works. Combining the foundational model with the fastest data processing engine on the market, Pathway enables enterprises to move beyond incremental optimization and toward truly contextualized, experience-driven intelligence.

Pathway is trusted by organizations such as NATO, La Poste, and Formula 1 racing teams.

The company is led by **co-founder and CEO Zuzanna Stamirowska**, a complexity scientist who created a team of AI pioneers:

- **CTO Jan Chorowski** was among the first people to apply Attention to speech and worked with Nobel laureate Geoff Hinton at Google Brain.
- **CSO Adrian Kosowski** is a leading computer scientist and quantum physicist who obtained his PhD at age 20 and co-founded SPOJ, one of the earliest popular competitive programming platforms.

The company is backed by leading investors and advisors, including Lukasz Kaiser, co-author of the Transformer ("the T" in ChatGPT) and a key researcher behind OpenAI's reasoning models. Pathway is headquartered in Palo Alto, California, with offices in Paris and Wroclaw.

Pathway invites you to join their open-source community on GitHub and leverage their tools to transform challenges into solutions

Pathway

LLM-App

The Challenge

You are given two things:

- A **complete long-form narrative** (a novel, 100k+ words)
- A **hypothetical backstory** for one of its central characters

The backstory is not part of the novel. It is newly written.
It is deliberately plausible.

Your task is to decide whether the proposed backstory is **consistent** with the story as a whole.

The goal is not to judge writing quality or check for small textual contradictions. Instead, you must determine whether the backstory respects the key constraints that have been established throughout the narrative.

What the System Is Expected to Demonstrate -

- **Consistency over time -** The system should check whether the proposed backstory fits with how characters and events develop later in the story.
- **Causal reasoning -** The system should determine whether later events still make sense given the earlier conditions introduced by the backstory.
- **Respect for narrative constraints -** Some explanations or coincidences don't fit a story, even if they don't directly contradict a sentence. The system should detect such mismatches.
- **Evidence based decisions -** Conclusions should be supported by signals drawn from multiple parts of the text, not from a single convenient passage.

Task Definition

Input
Each example contains:

1. Narrative

- Full text of a novel
- No summaries, no truncation

2. Hypothetical Backstory

- Character outline describing:
  - early-life events,
  - formative experiences,
  - beliefs, fears, ambitions,
  - assumptions about the world and its rules.

The backstory is intentionally underspecified in some places and overly confident in others.

Output
For each example, your system must produce:

1. Consistency Judgment

A binary label:
- Consistent (1)
- Contradict (0)

2. Comprehensive Evidence Rationale: Establishing Backstory Validity (Optional for Track B)

The Evidence Rationale is a critical component for substantiating or challenging the claimed backstory elements of the narrative. It must be meticulously constructed to ensure that all claims are rigorously tested against the primary textual source.

Structure and Requirements:

The Dossier must adhere to the following strict organizational principles to maintain academic rigor:

1. **Excerpts from the Primary Text:** Include direct, verbatim passages from the novel. These excerpts serve as the foundational textual evidence. The passages must be chosen for their direct relevance to a specific backstory claim, either explicitly supporting, contradicting, or providing crucial context for it.
2. **Explicit Linkage to Backstory Claims:** Every single textual excerpt must be unequivocally paired with a specific claim regarding the character's, world's, or plot's history (the 'backstory claim'). This linkage ensures that the evidence is always relevant and targeted. A single excerpt may constrain multiple claims, and a single claim may be addressed by multiple excerpts.
3. **Analysis of Constraint or Refutation:** For each linked pair (excerpt and claim), a concise yet thorough explanation must be provided.

The integrity of the analysis relies entirely on the proper linkage and support structure.

Tech Stack and Implementation

This challenge supports two parallel implementation tracks. Both address the same underlying task and dataset, but differ in ambition, technical depth, and evaluation philosophy.

Participants must choose one track to compete in. Submissions are evaluated only within their chosen track.

Track A: Systems Reasoning with NLP and Generative AI

This track is designed to encourage strong, well-engineered solutions using today's established NLP and GenAI techniques.
The focus here is correctness, robustness, and evidence-grounded reasoning, not architectural novelty.
Technical Requirements

All Track A submissions must use **Pathway's Python framework** in at least one meaningful part of the system pipeline.

Pathway may be used for:

- ingesting and managing the provided long-context narrative data,
- storing and indexing full novels and metadata,
- enabling retrieval over long documents using a vector store,
- connecting to external data sources (e.g. Google Drive, local folders, cloud storage),
- serving as a document store or orchestration layer for the reasoning pipeline.

Beyond this requirement, **all modeling choices are open**. Teams may use:

- transformer-based LLMs or Agentic Pipelines,
- classical NLP pipelines,
- hybrid symbolic - neural approaches,
- rerankers, classifiers, or custom heuristics.

Evaluation Focus

Submissions in Track A will be evaluated along the following dimensions:

- **Accuracy and robustness** on the core task (classification).
- **Novelty:** Thoughtful use of NLP or generative AI methods beyond basic or template-based pipelines. This may include custom scoring methods, step-by-step refinement, or small generative components used selectively to compare possible causes and effects, rather than using generation end-to-end.
- **Handling of Long Context:** Effectiveness in managing long narratives, such as chunking strategies, memory mechanisms, retrieval policies, or consistency checks that preserve global coherence.

The intent is to reward careful reasoning and novel NLP ideas, not straightforward applications of off-the-shelf RAG pipelines.

Track B: BDH-Driven Continuous Narrative Reasoning

Track B is intended for teams who want to explore BDH-inspired modeling more deeply and experiment beyond standard NLP pipelines.
In this track, participants are expected to engage meaningfully with ideas from Baby Dragon Hatchling (BDH), either by using the open-source implementation directly or by building components clearly motivated by its core principles.
The goal is to study how BDH-style mechanisms affect learning, representation, and classification over long narratives, rather than focusing solely on downstream explanation quality.

Technical Requirements

Track B submissions **must incorporate BDH** in one of the following ways:

- Using the open-source BDH architecture as part of the system.
- Pretraining or adapting BDH on task-relevant signals (e.g., narrative segments, event sequences, intermediate states).
- Using BDH to produce representations that are then fed into a classification or ranking head.
- Implementing reasoning components explicitly inspired by BDH principles, such as persistent internal state, selective or sparse updates, and incremental belief formation over time.

Participants are not required to train large-scale models or reproduce results from the original BDH paper. Small-scale experiments, partial implementations, and focused probes are fully acceptable.

Evaluation Focus

Submissions in Track B will be evaluated along the following dimensions:

- **Accuracy and robustness** on the core task (classification).
- **Pretraining** and Representation Learning using BDH
- **Clarity** in how BDH-style mechanisms influence representations or decisions, compared to standard transformer-based approaches.

Providing Evidence Rationale is optional, not required.
Submissions will not be penalized for focusing primarily on classification or prediction quality.

Dataset:

Dataset Link

The dataset consists of long-form narrative texts and associated hypothetical backstories, designed to evaluate global consistency and causal reasoning over extended contexts. Each example includes the complete text of a novel (100k+ words) provided in plain .txt format, hosted via a shared Google Drive link, ensuring no truncation or summarization of the source material. For every narrative, a separately written hypothetical backstory is provided for one central character. These backstories describe early-life events, beliefs, motivations, and assumptions, and are intentionally crafted to be plausible yet potentially inconsistent with the narrative.

Deliverables and Submission

Submit a single ZIP file named:
<TEAMNAME>_KDSH_2026.zip

The ZIP must contain the following:

1. Code (Reproducible)
- Runnable code that reproduces your results end-to-end
- Should read the provided inputs and generate predictions without manual steps

2. Report (Max 10 pages, excluding appendix)

A concise report describing:
- your overall approach,
- how you handle long context,
- how you distinguish causal signals from noise,
- key limitations or failure cases.

Clarity matters more than length.

3. Results File (CSV)
Submit a CSV file named results.csv with one row per test example.

Example Format:

- 1 = Backstory is consistent
- 0 = Backstory is inconsistent
- Rationale (optional but encouraged): Short explanation (1–2 lines).

Story ID	Prediction	Rationale
1	1	Earlier economic shock makes outcome necessary
2	0	Proposed backstory contradicts later actions
...	...	...

Participants are responsible for ensuring that their submission runs correctly in a clean environment.

Evaluation and Reproducibility

Submissions may be evaluated both on the reported outputs and by running the provided system. Inconsistencies between submitted predictions and system outputs may result in disqualification.

The evaluation prioritizes reasoning quality and robustness over raw performance. Submissions that demonstrate thoughtful design, clear reasoning, and honest discussion of limitations are favored.

Official Communication Channels

Whatsapp - Community 1, Community 2 (Join any one)
Discord - Channel

Resources

Pathway Framework:

- Core Engine
- LLM App templates
- Documentation
- Community Showcases
- Bootcamp

Connectors and Ingestion:

- Connectors overview
- Custom Python connectors
- Artificial data streams

LLM Integration:

- LLM xPack overview
- Pathway Vector Store docs

Tutorials:

- LangGraph agents cookbook

BDH Resources

1. Official Page:

- Main Repository
- Model architecture
- Training script

2. Paper (Key Sections):

- Full paper: HTML version
- **Section 2:** BDH architecture and local distributed graph dynamics
- **Section 3:** BDH-GPU tensor formulation
- **Section 6:** Interpretability findings, sparsity measurements, monosemantic synapses
- **Section 7:** Experimental validation and scaling laws
- **Appendix E:** Complete BDH-GPU code listing, alternative Implementations

3.Community Projects:

- HuggingFace Transformers compatible wrapper: Use AutoModel, AutoConfig, and .generate() with BDH. Supports both recurrent and parallel attention modes.
- MLX port for Apple Silicon
- Burn/Rust port
- Educational fork with visualizations: Recommended starting point for Path B of this track.
- Key files: bdh.py(model), boardpath.py(training/inference/config), utils/ (visualization).
- **Output examples:** combined_board_neuron.gif, combined_attention_sparsity.gif
- adamskrodzki/bdh: Dynamic vocabulary, stateful attention
- Git-Faisal/bdh
- GrahLnn/bdh