Agenda-based Narrative Extraction

Steering Pathfinding Algorithms with Large Language Models

Brian Keith Carolina Rojas Claudio Meneses Elizabeth Lam Angélica Flores
Ignacio Molina Joshua Leyton

Universidad Católica del Norte, Antofagasta, Chile
brian.keith@ucn.cl

Text2Story'26 @ ECIR 2026
Delft, The Netherlands, 29 March 2026

Outline

Why Narrative Extraction?

The Setting

Analysts face **large document collections** where events evolve over time. Reading everything is infeasible.

Goal: automatically extract **coherent storylines** from the corpus.

Applications

- News summarisation
- Intelligence analysis
- Disinformation research
- Investigative journalism
- Sensemaking workflows

Event-based Paradigm

- Each **document** $\equiv$ one event
- A **narrative** is a temporally ordered, thematically coherent sequence of document-events

Prior Work: Narrative Maps

How it works

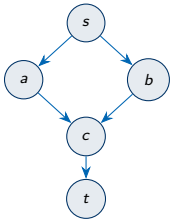
- Builds a **DAG** of documents
- Linear programming: balance coherence with **corpus coverage**
- Coverage constraints produce **multiple storylines** naturally
- Can capture user feedback through **semantic interaction**

Strengths

Rich interaction · Multiple storylines · Corpus coverage

Weakness

LP optimises **globally**: individual paths can have weak transitions. Coherence of each single storyline is not directly maximised.



multiple paths

Prior Work: Narrative Trails

Coherence Graph $G = (V, E, w)$

- Nodes = documents; edge weight = pairwise coherence $\theta(d_u, d_v)$
- Directed, acyclic (temporal constraints enforced)
- Sparsified via maximum spanning tree threshold

Maximum Capacity Path

$$\max_{P:\, s \to t} \; \min_{e \in P} w(e)$$

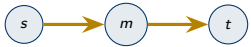
Solved via modified Dijkstra. Guarantees every transition meets a minimum quality threshold.

Strength

Maximises the **weakest link**, producing highly coherent individual storylines.

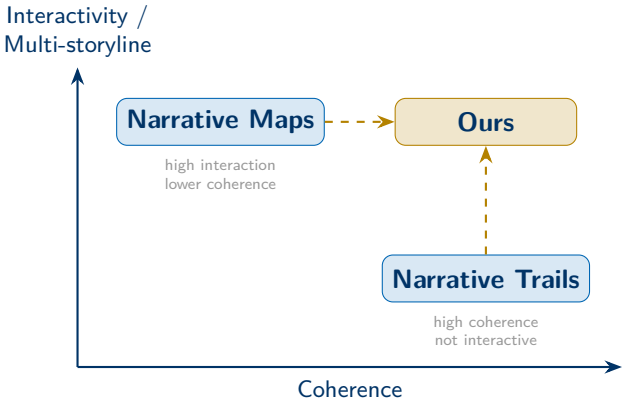
Weaknesses

- Single, **deterministic** output
- No user guidance possible
- Cannot generate **alternative framings**



single optimal path

The Design Space Gap & Research Questions



RQ1

Can LLM-guided steering **align** narratives with user-specified agendas whilst maintaining coherence?

RQ2

What is the **trade-off** between agenda alignment and narrative coherence?

Our goal

Preserve Trails' coherence whilst adding Maps' interactivity via **natural language agendas**.

What Is an Agenda?

Definition

An **agenda** A is a natural language string describing the **perspective** the user wishes to explore in the corpus.

It expresses a *framing*, not a retrieval query.

Key Properties

1. **User-specified:** any natural language definition
2. **Perspective-bearing:** not just a topic, but a framing that we wish to emphasize

Augmenting IR with LLM guidance

This is a form of IR augmented with LLM-guided narrative framing. Rather than matching documents to the agenda as a query, we **steer a coherence-preserving path** towards the desired perspective.

Building the Coherence Graph

Construction pipeline:

1. Embed documents with **Sentence-BERT**
2. Reduce dimensionality via **UMAP**
   (avoids curse of dimensionality for HDBSCAN; enables direct visualisation)
3. **Soft-cluster** with HDBSCAN: each document d receives topic distribution p_d
4. Compute **coherence** for each event pair
5. **Sparsify**: retain edges above a threshold derived from the maximum spanning tree

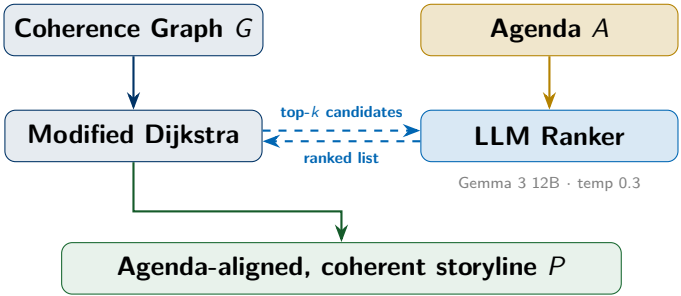
Coherence Function

$$\theta(d_u, d_v) = \sqrt{S(\hat{z}_u, \hat{z}_v) \cdot T(p_u, p_v)}$$

- S = angular similarity (2D UMAP projections)
- $T = 1 - \mathrm{JSD}(p_u, p_v)$
  topical similarity via Jensen-Shannon divergence

Geometric mean: requires *both* spatial proximity and topical overlap to be high.

Agenda-Driven Pathfinding: System Overview



Running the algorithm with different agendas on the *same* corpus produces different coherent storylines.

LLM Ranking: The Prompts

Direct Ranking Prompt (default)

```
You are helping build a narrative aligned
with: {agenda}
Context: {titles of articles so far}
Current article: {current title}
Target article: {target title}
Rank ALL from BEST to WORST:
{1: title_A, 2: title_B, ...}
Respond: {"ranking": [3,1,2]}
```

Chain-of-Thought Variant (ablation)

```
Step 1 - Perspective alignment:
Does each option support, contradict, or remain
neutral?
Step 2 - Path to destination:
Which options help reach the target?
Step 3 - Ranking:
Order all options from best to worst.
Respond: {"reasoning":"...", "ranking":[3,1,2]}
```

Prompt Design

The prompts could be further improved. However, we note that the goal was not finding the optimal prompt, but showing that steering helps to align the extracted narratives.

Baseline Methods

Maximum Capacity (agenda-agnostic)

The unmodified Narrative Trails algorithm.

Equivalent to the agenda-based method with $k = 1$: no LLM call and pure coherence optimisation.

Serves as the **coherence upper bound** and agenda-agnostic reference.

Keyword Matching (TF-IDF)

Score candidates by:

$$\text{score} = (1 - \alpha) \cdot \theta + \alpha \cdot \text{tfidf-sim}$$

with $\alpha = 0.5$. Tests whether **lexical-overlap alone** can achieve agenda alignment without semantic reasoning.

Method	Agenda-aware	Semantic understanding	Coherence-preserving
Max Capacity	×	×	✓ (best)
Keyword	✓	×	✓
LLM-Driven	✓	✓	✓

Dataset: 2021 Cuban Protests Corpus

Corpus Details

- **418 news articles** (Dec 2020 – Oct 2021)
- Predominantly **US news sources**
- Topics covered: protests, government crackdown, Cuban-American diaspora, US foreign policy, international coverage

Endpoint Selection

Source–target pairs that are:

- Temporally ordered
- Connected in the coherence graph
- Prioritised by **UMAP distance**

Sample sizes (power analysis):

- Main evaluation: **64 pairs** ($d = 0.5$)
- Ablations: 26 / 21 / 17 pairs

Six Agendas in Three Categories

Simple (literal)

Keywords appear **verbatim** in the corpus.

Freedom Uprising
"Cubans demanding freedom from communist rule"

Diaspora Solidarity
"Cuban-Americans rallying to support protesters in Cuba"

Semantic (inference)

Requires reasoning **beyond literal overlap**.

Regime Crackdown
"Cuban regime violently suppressing protesters"

Gov. Censorship
"Cuban government controlling information through internet restrictions"

Counter (negative control)

Contradict the corpus; test whether steering can fabricate.

Protests Failing
"Cuban protests losing momentum"

Regime Popular
"Cuban government maintaining popular support"

Simple: keyword matching expected competitive. **Semantic**: LLM steering expected to excel. **Counter**: tests whether fabrication is possible.

Total evaluations in the main experiment: $64 \times 6 \times 3 = \mathbf{1{,}152}$ evaluations

Evaluation: Metrics for LLM Judges

Coherence (1–10 scale)

Four sub-dimensions:

- Logical flow
- Thematic consistency
- Temporal coherence
- Narrative completeness

Score 5 = adequate · 7 = good · 10 = exceptional

Alignment (1–10 scale)

Five sub-dimensions:

- Agenda support
- Persuasiveness
- Evidence strength
- Narrative direction
- Bias effectiveness

3–5 = surface relevance only; 7+ = true argumentative alignment

Input to judges: full concatenated text of all articles in the extracted path, in sequence order.

Evaluation: LLM Judges

Two independent judges

Claude Opus 4.5 + GPT 5.1; report the *mean* of both scores.

Mitigates individual model biases; validated as a proxy for human evaluation of narrative coherence.

Inter-rater Reliability

	Pearson r	Spearman ρ
Alignment	**0.88**	**0.93**
Coherence	0.53	0.50

Strong agreement on alignment

$r = 0.88$, $\rho = 0.93$: both judges largely agree on which narratives are persuasively agenda-aligned.

Moderate agreement on coherence

$r = 0.53$: coherence is harder to judge consistently. Human evaluation is planned as future work to complement LLM-based assessment.

Main Results: Overall Scores

Method	Coherence	Alignment	Path Length
Maximum Capacity	**6.33**	4.48	10.8
Keyword Matching	6.28	4.64	11.8
LLM-Driven (k=5)	6.19	**4.80**	11.7

Best overall alignment

LLM-Driven achieves highest alignment. Coherence cost: only **0.14 pts (2.2%)** compared to Maximum Capacity.

Path length

Agenda-driven methods produce slightly longer paths ($\tilde{1}1.75$ vs. 10.80 articles), showing more exploration of topical regions.

Results by Agenda Type

Method	Overall	Literal	Semantic	Counter
Maximum Capacity	4.48	—	—	—
Keyword Matching	4.64	**6.91**	4.78	2.24
LLM-Driven (k=5)	**4.80**	6.65	**5.25**	2.49

Literal agendas

Keyword wins (6.91 vs. 6.65). LLM overhead not justified when vocabulary is shared.

Semantic agendas

LLM wins: **+9.9%** vs. keyword ($p = 0.017$). Semantic inference matters.

Counter agendas

All methods: 2.24–2.49. **No fabrication.**

Counter-Agenda Validation

All methods score uniformly low

Agenda	Mean	Spread
Protests Failing	≈2.5	0.25 pts
Regime Popular	1.94 (lowest)	0.25 pts
Simple agendas	≈6.8	0.26 pts
Semantic agendas	≈5.0	0.47 pts

All methods converge to the same low floor.

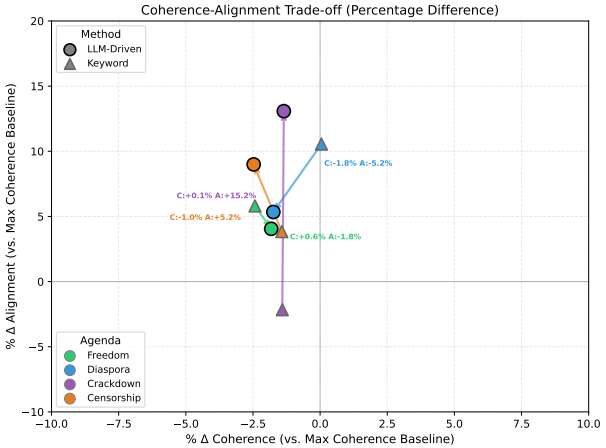
This is a *feature*

The corpus covers **growing protests and regime opposition**.

No document selection can reverse this: the data simply does not contain pro-regime narratives.

Steering amplifies; it does not fabricate.

Coherence vs. Alignment Trade-off



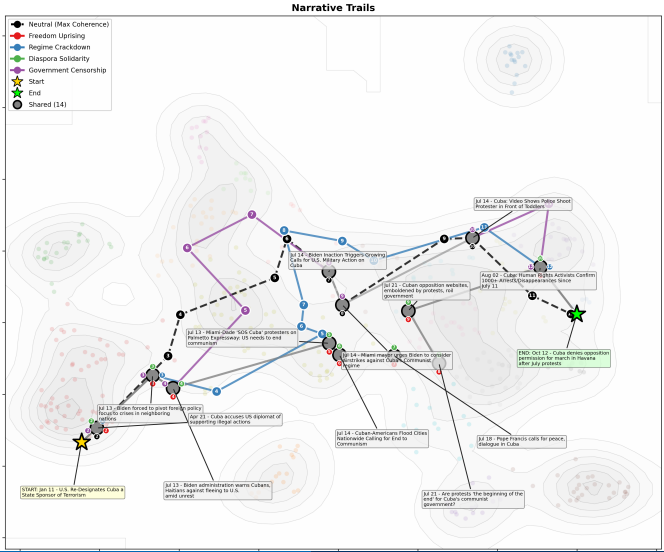
Each point: one method × agenda, expressed as % difference from the Maximum Capacity baseline.

- **Crackdown (LLM):** approx. +15% alignment, −1% coherence
- **Diaspora (Keyword):** approx. +10% alignment
- Pearson $r \approx 0.10$ across all 1,152 evaluations

Conclusion

The trade-off is **not fundamental**. High alignment does not require sacrificing coherence.

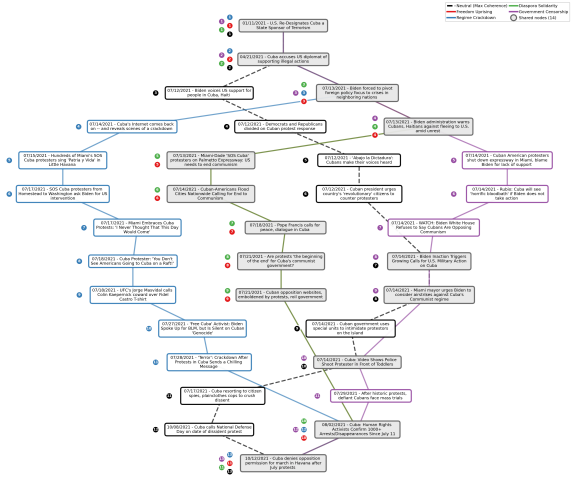
Visualisation: Paths Through UMAP Space



What to notice

- Each colour = one agenda-driven path
- Paths **diverge through different topical regions**
- *Regime Crackdown*: government-response clusters
- *Freedom Uprising*: protest-coverage clusters
- Jaccard similarity (baseline vs. agenda paths): **0.25–0.50**

Visualisation: Narrative Map Structure



Structure

- All five paths flow **top to bottom**
- **14 shared nodes** highlighted
- Distinct routes despite identical endpoints
- All paths within **2.2%** of baseline coherence

Multiple storylines via agendas

4 agendas and the baseline produce 5 coherent narratives through the same corpus.

Sensitivity Analysis

Hyperparameter sweep

Parameter	Range
Candidate pool k	1–10
Temperature	0.1–1.0
Model size	270M–12B

- **Coherence** stable across all settings
- Vague agendas perform as well as or better than specific ones
  - ▷ Over-specification may constrain the LLM's ability to recognise relevant articles

Chain-of-Thought: Biggest Effect

	Direct	CoT
Alignment	4.04	**5.07**
Coherence	6.15	6.15
Time/narr.	32s	**161s**
Jaccard	0.58	

+25.7% alignment ($p = 0.017$) at **5**× cost.

Jaccard of 0.58: meaningfully different document selections between the two prompting modes.

When Does LLM Steering Help?

Use LLM steering when

- Agenda requires **semantic inference**: vocabulary gap between agenda and corpus
- Latency is not a hard constraint

Prefer keyword matching when

- Agenda keywords appear **literally** in the corpus: there is no need for semantic inference
- Real-time or interactive latency is required

Practical recommendation

Classify your agenda first. Semantic agendas benefit from LLM steering; literal agendas do not justify the cost. A **hybrid** strategy (using keyword matching to pre-filter and the LLM only when needed) is a natural next step.

Limitations

Evaluation scope

- Single corpus (2021 Cuban protests): generalisability to other domains is unknown
- Only 6 agendas in 3 categories
- Ablations on 17–26 endpoint pairs: directional guidance only; limited statistical power

Methodology

- LLM judges may carry systematic biases not captured by inter-rater agreement
- **Human evaluation** planned: to validate alignment with analyst judgements in downstream sensemaking tasks
- LLM steering adds latency; this may be hardware-dependent rather than fundamental

Ethical Considerations

Potential misuse

Agenda-driven extraction could be used to construct **misleading narratives** from selectively chosen documents, for instance in disinformation campaigns designed to support a false conclusion.

Mitigating factors

1. Makes framing **explicit and studiable** rather than hidden
2. Cannot fabricate content absent from the source data (counter-agenda results confirm this)
3. **Understanding** steering is a prerequisite for *detecting* it in real-world systems

Conclusion & Future Work

Summary

Agenda-based narrative extraction bridges Narrative Maps and Narrative Trails, adding:

1. **Interactivity** via natural language agendas
2. **Multiple storylines** via agenda variation

Key results:

- **+9.9%** alignment on semantic agendas ($p = 0.017$)
- **+13.3%** on Regime Crackdown ($p = 0.037$)
- Only **2.2%** coherence cost
- Counter-agenda scores 2.2–2.5: no fabrication

Future Work

- Multi-domain evaluation
- Human evaluation of narrative quality
- Hybrid LLM and keyword steering
- Agenda auto-generation from user intent
- Detecting steered narratives in the wild
- Latency reduction for interactive use

Thank You!

Agenda-based Narrative Extraction:

Steering Pathfinding Algorithms with Large Language Models

Brian Keith, Carolina Rojas, Claudio Meneses, Elizabeth Lam, Angélica Flores, Ignacio Molina, Joshua Leyton

Universidad Católica del Norte, Antofagasta, Chile `brian.keith@ucn.cl`

Acknowledgements

This research is funded by ANID FONDEF ID25I10072 (*Narrative Panopticon: Intelligent Platform for Mapping and Monitoring Information Narratives from Multi-Source Data Streams*). Supported by ANID FONDECYT 11250039 (*Interactive Narrative Analytics*) and Project 202311010033-VRIDT-UCN. We thank the *coreDevX* team for supporting the Narrative Panopticon project.

NARRATIVE PANOPTICON

UNIVERSIDAD CATÓLICA DEL NORTE CHILE

coreDevX®

Agencia Nacional de Investigación y Desarrollo

Ministerio de Ciencia, Tecnología, Conocimiento e Innovación

Gobierno de Chile