Can Large Language Models Facilitate Qualitative Political Narrative Analysis?

Luke Stephens, Clare Llewellyn, Lauren Rogers, Constantine Kyritsopoulos, Feiteng Long, Arman Prangere, Peyton Snyder, Laura Cram

Neuropolitics Research Labs, University of Edinburgh
Old College, South Bridge, Edinburgh, EH8 9YL, UK;
Luke.Stephens@ed.ac.uk, Clare.Llewellyn@ed.ac.uk, L.K.Rogers@ed.ac.uk,
C.G.Kyritsopoulos@ed.ac.uk, Feiteng.Long@ed.ac.uk, A.T.P.Prangere@ed.ac.uk,
P.Snyder@ed.ac.uk, Laura.Cram@ed.ac.uk

Abstract

This study evaluates whether Large Language Models (LLMs) can facilitate qualitative political narrative analysis by comparing outputs from four models—Mistral, Llama, ChatGPT-4o, and DeepSeek—against narrative analyses written by expert scholars. Using European Union State of the Union speeches (2010–2023), we examine migration and solidarity narratives through semantic and lexical similarity metrics alongside systematic validation. The narrative scholars demonstrate strong semantic alignment despite differences in wording, establishing a benchmark for interpretive consistency. Across both topics, the models produce lexical and semantic similarity scores that are broadly comparable to those observed between the scholars themselves, with differences at these levels often marginal. However, similarity metrics do not provide the full picture. Validation reveals model-specific weaknesses that are not captured by lexical or semantic alignment alone, including factual errors, over-structural abstraction, and difficulty engaging less salient narrative threads. These findings demonstrate that LLMs can produce narratives that align closely with human outputs in semantic and lexical similarity, yet these measures alone are insufficient to assess interpretive quality.

Keywords: : Large Language Models, Qualitative Analysis, Narrative Analysis, Political Communication, Computational Methods

1. Introduction

Narratives serve as both vital strategic tools and foundational structures of global politics. Narrative research has evolved beyond its origins in literary theory to become a central analytical framework across the social sciences, offering schemas through which individuals and societies understand history, identity, and behaviour. They bring actors, objects, and events into temporal and spatial connections. On a strategic level, narratives represent tools used by political actors to achieve their ambitions, shaping responses to events or reaching desired policy outcomes. More fundamentally, narratives actively produce the meanings through which the social world is understood. Narrative research within global political studies is predominantly qualitative and interpretive, reflecting the context-specific and constructivist assumptions shaping much of political narrative scholarship.

Alongside the growing focus on narrative scholarship in the social sciences, computational approaches have increasingly engaged with the concept of narratives. Scholars have applied Natural Language Processing (NLP) tools to the study of narratives, recognising their temporal and relational properties as analytically distinctive forms of discourse. Computational work focuses on extracting discrete narrative components, such as named entity recognition, plot archaeology, character network analysis, and event identification, (Spiliopoulou et al., 2017)(Rahimtoroghi et al., 2017) (Dekker et al., 2019) rather than analysing narratives as coherent stories. Useful for systematic extraction, such structural approaches prioritise the identification of discrete components over the broader connections between them that produce coherent and complete narratives, limiting their usefulness for qualitative and interpretive research.

The emergence computational social science approaches and the use of Large Language Models (LLMs), which can both identify narrative elements like traditional NLP tools and generate narrative text. This offers new for opportunities combining computational methods within qualitative and interpretive narrative analysis.

We present a methodology for narrative analysis using LLMs as an assistant in this qualitative research, assessing whether LLMs can replicate core processes within interpretive analysis and evaluating those outputs through structured human validation. We use the European Union State of the Union speeches (SOTEU) — annual addresses delivered since 2010 in which the Commission President reflects on the previous year and sets out priorities for the next. We compare narrative summaries pro-

Proceedings of The 3rd Workshop on Natural Language Processing for Political Sciences (PoliticalNLP 2026) @ LREC 2026, pages 261–271
16 May 2026, Palma, Mallorca (Spain).

duced by two human narrative scholars focusing on two topics (migration and solidarity), with four language models — DeepSeek, Llama, ChatGPT-4o, and Mistral. Model outputs were compared with the human narratives using lexical and semantic similarity measures, and systematic human validation assessing structural, factual, and narrative elements. The findings suggest LLMs can produce narratively coherent accounts of political texts that reflect similar interpretive readings to those of human scholars, though systematic human validation remains essential to identify factually inaccurate and catastrophic errors. These results provide the foundations for a Human-LLM Co-Analysis approach that allows qualitative researchers to engage with larger corpora.

2. Literature Review

2.1. Narrative Analysis in Political Science

The narrative turn in the social sciences reflects narratology's evolution from fictional storytelling toward understanding social reality. Narratives impose order on otherwise disconnected events through emplotment—constructing causal links rather than mere chronology. Ricoeur (1984) argues that storytelling is fundamentally interpretative, imposing patterns and explanations on experiences that would otherwise appear disordered, White (1987) demonstrates how narratives structure historical understanding. In political life, narratives function both strategically and as the basis of a shared social reality. Political actors deploy narratives to frame events, establish legitimacy, and influence public opinion, shaping discursive environments to advance particular goals (Chaban et al., 2019). The persuasive force of a narrative depends on resonance with cultural myths, collective memories, and emotionally salient identities. Political actors can create narratives, but they are constrained by broader historical and social storylines that define plausibility and persuasion. Narratives can be strategic tools (Chaban et al., 2019; Colley and {van Noort}, 2022), structures for shaping identity (Somers, 1994; Steele, 2008), and the international order itself (Hagström and Gustafsson, 2019; Homolar and Turner, 2024).

The understanding of narratives as meaning-making practices underpins qualitative narrative analysis. Built upon a constructivist ontology, such research views social reality as co-created through language and storytelling rather than as fixed or objective. Narratives link events, actors, and motivations into coherent meaning structures, shaping how political worlds are experienced and understood. Epistemologically, narrative inquiry is grounded in interpretivism: knowledge is context-bound, and meaning cannot be separated from interpretation. Within this framework, narratives are analysed as coherent structures in which setting, characterisation, and emplotment work together to shape how political realities are understood and experienced.

2.2. Computational Approaches to Narrative Analysis

Computational methods have often engaged from a different analytical starting point. Over the past decade, computational approaches expanded alongside increasingly sophisticated Natural Language Processing techniques, focusing on automating the detection and classification of narrative components, identifying fragmented storylines, tracing temporal arcs, and mapping rhetorical framing at scale. This work centres largely on structural elements such as events, characters, and settings, treating narratives as data to be segmented and categorised. Event detection was prominent, from Spiliopoulou et al. (2017)'s frame semantic parsing to Rehm et al. (2017)'s Movement Action Events linking actors, events, and settings. Other studies traced thematic evolution, including Schmidt (2015)'s analysis of television synopses and Hussain et al. (2018)'s work on migration narratives. Dekker et al. (2019) used Named Entity Recognition to extract characters and roles, Rahimtoroghi et al. (2017)'s DesireDB tracked character goals, and Hu and Walker (2017) modelled causal relations. Useful for systematic extraction, such structural approaches prioritise the identification of discrete components over the broader connections between them that produce coherent and complete narratives.

The emergence of LLMs extended these computational approaches, engaging with narrative components and as a coding tasks. Applications are focused on component detection: Michelmann et al. (2023) found ChatGPT-3's event detection comparable to human annotators, Stammbach et al. (2022) extracted archetypal roles, and Sun et al. (2024) reported improved causal relation identification. Piper and Bagga (2025) demonstrated that LLMs capture narrative dimensions such as point of view and sequencing, though they struggled with abstraction. LLM's have been used to replicate or assisting human coding schemes, with Bano et al. (2023) reporting partial alignment, and Dunivin (2025) proposed "scaling hermeneutics" where human codebook development precedes machine application.

Benoit et al. (2025) use LLM in political text generation, developing pipelines generating manifesto summaries, and validating outputs against

Year	Narrator	Word Count
2010	José Manuel Barroso	4384
2011	José Manuel Barroso	5046
2012	José Manuel Barroso	6077
2013	José Manuel Barroso	5628
2015	Jean-Claude Juncker	10027
2016	Jean-Claude Juncker	6049
2017	Jean-Claude Juncker	6245
2018	Jean-Claude Juncker	5237
2020	Ursula von der Leyen	8249
2021	Ursula von der Leyen	6512
2022	Ursula von der Leyen	5795
2023	Ursula von der Leyen	6797
Total		**76046**

Table 1: EU State of the Union Addresses: Narrators and Word Counts

expert surveys. Jenner et al. (2025) compared human-written and LLM-generated qualitative narrative analyses of personal narratives, showing thematic convergence under structured prompting but documenting hallucination risks requiring human supervision and checks.Treynor and McCoy (2025) demonstrated that LLMs generate narrative fragments with varying intensity levels, with users reliably perceiving these differences as intended. Maisto (2025)'s linguistic analysis of automatically-generated role-playing game sessions revealed that LLM language exhibits patterns distinct from both human speech and written text, particularly in textual cohesion and narrative structure.

We extend these approaches by defining a methodological framework for qualitative political narrative analysis, benchmarking multiple models against human scholars using lexical and semantic metrics, and systematic human validation. This redefines evaluation and validation as a measure of the interpretive and narrative quality of outputs.

3. Methodology

3.0.1. Data Selection

This study focuses on the annual EU State of the Union addresses delivered by the President of the European Commission. Since 2010, these speeches have provided a central platform for addressing Members of the European Parliament, European political leaders and the wider European public. Introduced following the Lisbon Treaty, the State of the Union is framed by the European Commission as an opportunity to review the previous year and outline priorities for the year ahead.

This analysis concentrates on two recurring narrative topics: migration and solidarity. Migration has represented a sustained political challenge for the European Union, particularly following the 2015 refugee crisis, and is frequently mobilised in domestic contestation of the European project. European solidarity, by contrast, is a constitutive theme within EU discourse, reflecting expectations of burden-sharing, shared values, and institutional cooperation. The dataset comprises 13 speeches delivered by three Commission Presidents, totalling approximately 75,000 words (Table 1). All speeches were delivered in English. Although modest by computational standards, the corpus enables systematic comparison between human and model-generated narratives while remaining manageable for detailed interpretive engagement alongside quantitative similarity analysis.

3.0.2. Research Design

The study proceeded in four stages. First, two human narrative scholars produced narrative analyses for each speech, writing two 300-word texts per address — one on migration and one on solidarity, as well as a single 600-word analysis examining narratives on the topics across the full corpus of speeches over time. Both scholars were given the same task and produced their analyses independently, without access to each other's outputs. Second, four LLMs were given the same tasks and instructed to generate equivalent analyses. For the address specific narrative the prompts contained the address for the specified year and either: **Migration:** *"You are a narrative scholar tasked with writing a narrative analysis of up to 300 words for the following document. You are analysing narratives of migration, and this topic can include migration to Europe, migration within Europe, and migration out of Europe."* **Solidarity:** *"You are a narrative scholar tasked with writing a narrative analysis of up to 300 words for the following document on the topic of European solidarity. European solidarity involves working together to create an integrated union shaped by mutual support, shared values, history, heritage and ideals. Citizens, national governments of the members states and EU institutions work collaboratively and adhere to common policies, based on: consensus on European interests; collective European identity; empathy; trust; social fairness and the avoidance of inequalities; and underpinned by the rule of law."* The prompts were intentionally limited in their specification of what constitutes a narrative analysis, avoiding an explicit definition in order to assess the assumed knowledge and default interpretive tendencies of each model.

For the narratives on the topics across the full corpus of speeches over time the models were pro-

vided with the summaries they had created for each specific address, they were asked for a 600 word narrative analysis, they were told they were producing a summary from summaries. Third, model outputs were compared with the human narratives using lexical and semantic similarity measures, providing a structured assessment of textual alignment across surface and semantic levels. Fourth, all model-generated narratives were subjected to systematic human validation assessing structural, factual, and narrative dimensions, alongside an overall assessment of the quality of the LLM narrative outputs.

3.0.3. Models and Evaluation Metrics

The study evaluated four large language models: three small versions downloaded and used locally, Llama3.2:3B, Deepseek:r1, and Mistral:7b(0.3), from the open-source Ollama, and ChatGPT-4o via an API provided by the University of Edinburgh (ELM). API access to ChatGPT-4o was removed during this project, as OpenAI indicated this model was to be retired soon. This meant that we instead had to use ChatGPT4-turbo for the 600-word full corpus study, which was conducted in a later phase of the project. These models were selected to capture variation across widely used systems, smaller models that can run on standard machines, and a larger model. The change in GPT models available for our use and our necessary shift from GPT-4o to GPT-turbo during this research highlights the difficulties in using larger models provided by services over which we had no control. This reflects tools that are realistically available to researchers working outside large-scale computational environments. The temperature (amount of variability allowed in the model) was set to zero for all models. The models were tested over 100 runs and all produced the same output each time.

Similarity between human and model-generated narratives is evaluated using both lexical and semantic measures. Lexical similarity was calculated using Jaccard similarity (spaCy) to examine overlap in wording and surface-level expression. Semantic similarity was computed using Word2Vec (spaCy), and Doc2Vec (gensim, stsb-roberta-large), enabling comparison sentence, and document levels. This was generated using python scripts and libraries.

In addition to similarity metrics, all model outputs were subjected to systematic human validation across three categories. **Structural checks** assessed basic formal requirements: correct language and adherence to the word count. **Factual accuracy** examined whether named events, actors, and policies were real, relevant to the speech, and if key instances had been identified. **Narrative elements** assessed the interpretive dimensions of the analysis, including tone, event ordering, the contextualisation of events through the perspective of the narrating actor, the construction of relationships between events and broader symbols or values, and the presence of temporal connections between actors and events across the narrative. **Narrative quality** assessed the overall quality of the narrative analysis. Validators responded yes or no to each item before providing an overall quality rating on a 0–5 scale, where 0 indicated substantial factual or interpretive error, and 5 indicated a factual and accurate account. This ensured that alignment scores were complemented by systematic checks on factual grounding and narrative form.

4. Results

Analysis begins with the human–human baseline, which establishes the reference point for alignment across both migration and solidarity narratives. Semantic similarity between the two narrative scholars is consistently high (Table 3). These results indicate strong convergence in how events, actors, and narrative structure are interpreted, even across independently produced texts.

Lexical similarity is notably lower. The Jaccard scores show an average of 0.16, reflecting variation in phrasing and word choice despite shared semantic framing. Together, these results confirm that interpretive agreement does not require lexical replication. High semantic alignment combined with lexical divergence establishes an important benchmark for evaluating model performance.

4.0.1. Model Performance on Similarity Metrics

The four LLMs show consistently strong semantic alignment with the narrative scholars, though with variation across systems, indicating close alignment at sentence and document levels. In several instances, model–human semantic similarity approaches the range observed between the two scholars themselves. DeepSeek performs less consistently, displaying greater variability.

Lexical similarity remains lower across all systems, consistently below the human–human baseline. Mistral and ChatGPT-4o generally occupy the upper end of this range. However, these lexical differences mirror the broader pattern seen in the human baseline: semantic alignment remains comparatively strong even where surface-level wording diverges. Across the corpus, semantic similarity proves more stable and informative than lexical overlap.

Code	Question	Scale
Structural Checks		
Language	Are the narratives in the correct language?	Yes/No
Word_count	Are the narratives within the stated word count?	Yes/No
Factual Accuracy		
Real_event	Are the named events in the output real and related to the speech?	Yes/No
Key_event	Are the key events identified?	Yes/No
Real_actor	Are the actors in the output real and related to the speech?	Yes/No
Key_actor	Are the key actors identified?	Yes/No
Real_policy	Are the named policies in the output real and related to the speech?	Yes/No
Key_policy	Are the key policies identified?	Yes/No
Narrative Elements		
Tone_speech	Does the output accurately depict the tone of the speech?	Yes/No
Order_event	Does the output reflect events occurring in the correct order?	Yes/No
Narrating_actor	Does the output accurately reflect the institutional dynamics/structure of the narrating actor?	Yes/No
Values_event	Does the output construct a relationship between events to broader symbols, meanings, or values?	Yes/No
Perspective_narrator	Does the output contextualise events through the perspective of the narrating actor?	Yes/No
Connect_event	Does the output connect similar and related events to each other?	Yes/No
Semantic_setting	Does the output construct a semantic setting?	Yes/No
Temporal_emplotment	Does the output contain emplotment between actors and events in a temporal order?	Yes/No
Overall Quality		
Narrative_quality	Overall, how do you rate the quality of this narrative analysis?	0–5

Table 2: Validation Coding Scheme

Model	Migration						Solidarity					
	Jaccard		Word2Vec		Doc2Vec		Jaccard		Word2Vec		Doc2Vec	
	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
Human	0.163	0.018	0.874	0.038	0.995	0.005	0.175	0.019	0.903	0.036	0.998	0.001
DS	0.115	0.025	0.807	0.075	0.990	0.015	0.111	0.031	0.757	0.168	0.981	0.051
Llama	0.121	0.020	0.859	0.044	0.993	0.011	0.139	0.016	0.868	0.057	0.996	0.002
Mistral	0.139	0.022	0.898	0.033	0.992	0.011	0.148	0.015	0.903	0.033	0.997	0.001
GPT-4o	0.134	0.020	0.872	0.041	0.993	0.011	0.126	0.016	0.859	0.050	0.997	0.001

Table 3: Lexical and Semantic Similarity Scores by Model and Topic by Year

4.0.2. Topic and Temporal Variation

Patterns of alignment vary by topic. Narratives focused on solidarity consistently produce higher similarity scores than migration, apart from the Deepseek Model. Migration narratives show greater variability. We found that temporal variation further shapes alignment. The years 2015 and 2018 stand out as periods of peak model–human similarity. By contrast, 2020 and 2023 record lower alignment across most models. Importantly, similar fluctuations are visible in the human–human baseline. Reduced similarity in certain years therefore reflects discursive complexity rather than model failure alone.

4.0.3. Model Performance Across Corpus

Lexical and semantic performance of models across the whole corpus is equivalent to or higher than the individual year-by-year address summaries. There does not seem to be a drop off in scores when the models create a single summary based on summaries rather than the actual speech.

Model	Migration						Solidarity					
	Jaccard		Word2Vec		Doc2Vec		Jaccard		Word2Vec		Doc2Vec	
	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
Human	0.174	N/A	0.930	N/A	0.999	N/A	0.183	N/A	0.947	N/A	0.998	N/A
DS	0.135	0.009	0.872	0.018	0.998	0.000	0.134	0.001	0.900	0.028	0.997	0.001
Llama	0.110	0.003	0.896	0.009	0.998	0.000	0.122	0.002	0.923	0.009	0.998	0.001
Mistral	0.131	0.001	0.921	0.003	0.998	0.001	0.143	0.012	0.863	0.006	0.997	0.000
GPT-4o	0.128	0.017	0.923	0.013	0.999	0.000	0.148	0.018	0.894	0.016	0.998	0.001

Table 4: Lexical and Semantic Similarity Scores by Model and Topic Across the corpus (600-Word Analyses)

4.0.4. Validation

Similarity metrics capture alignment patterns, but do not fully assess structural and factual accuracy or narrative coherence and interpretation. Following the similarity analysis, all model-generated narratives were subjected to systematic human validation.

Structurally: all models produced narratives in english consistently except for DeepSeek which occasionally provided narratives that were not. A shared issue across all models was a failure to adhere to the 300-word limit. In both the migration and solidarity narratives, the majority of models produced outputs that exceeded the specified length. This was particularly pronounced in migration texts.

Factually: in the migration narratives events, actors and policy were identified by all models except DeepSeek. Llama performs less well at identifying the key events, actors and policy, although it does this well in the solidarity topic.

Narratively: DeepSeek does not perform well. Llama is looses accuracy when identifying connected events and temporal emplotment.

Narrative Quality: DeepSeek and Llama are generally inconsistent across the topics, Llama performing more poorly on migration and DeepSeek on solidarity. Both Mistral and ChatGPT-4o perform well.

Taken together, validation demonstrates that high semantic and lexical alignment does not guarantee factual or structural accuracy. Human evaluation remains essential for identifying errors that similarity measures may overlook.

4.1. 600-Word Analyses

Similarity scores for the 600-word analyses were generally strong across models, clustering closely around the human baseline across both migration and solidarity. This represents a broadly comparable pattern to the 300-word analyses. Validation results revealed sharper differentiation between models.

Structurally: all models received perfect scores.

Factually: most models correctly identify real events, actors, and policies, though Llama scored zero on key events and key actors.

Narratively: DeepSeek and GPT-4-turbo performed strongest, scoring well across tone, event ordering, and temporal emplotment, reflected in their higher quality scores. Mistral performed well across most categories but struggled on tone and semantic setting. Llama scored poorly across narrative elements, particularly on tone, event ordering, and semantic setting.

Narrative Quality: DeepSeek's perfect solidarity quality score stands in notable contrast to its performance on the 300-word analyses, where factual errors were most severe.

5. Discussion

The findings have important implications for the use of LLMs in interpretive qualitative narrative analysis. Similarity scores indicate that model outputs broadly reflect the semantic and lexical content of human analyses. Their significance lies in what systematic human validation reveals, namely that such alignment must be accompanied by validation to assess factual accuracy and interpretive coherence. Semantic and lexical alignment between models and the two narrative scholars across migration and solidarity indicates that LLMs can produce accounts that broadly correspond to human analyses in wording and semantic tone, though this varies across models. Although migration narratives generated lower scores, this pattern mirrors the human–human baseline, and variation is expected in interpretive research where no single reading is definitive.

The limits of similarity measures are visible in validation. Structurally, all models in the 300-word analyses exceeded the word limit, and in one case DeepSeek produced a response entirely in Mandarin. The 600-word analyses showed the inverse pattern, with models typically producing narratives of around 450–500 words and not making full use of the analytical space. Llama presented a more fundamental failure, producing a narrative of the 2022 speech only despite being tasked with analysing

	Structural Checks				Factual Accuracy			
LLM	Language		Word_count		Real_events		Key_events	
	mean	std	mean	std	mean	std	mean	std
DS	1.000	0.000	0.083	0.282	0.667	0.482	0.625	0.495
Llama	1.000	0.000	0.167	0.381	0.958	0.204	0.458	0.509
Mistral	1.000	0.000	0.083	0.282	1.000	0.000	0.875	0.338
GPT-4o	1.000	0.000	0.917	0.282	1.000	0.000	0.917	0.282
	Factual Accuracy							
	Real_actors		Key_actors		Real_policy		Key_policy	
	mean	std	mean	std	mean	std	mean	std
DS	0.625	0.495	0.667	0.482	0.667	0.482	0.708	0.464
Llama	0.917	0.282	0.750	0.442	0.917	0.282	0.708	0.464
Mistral	1.000	0.000	1.000	0.000	1.000	0.000	0.958	0.204
GPT-4o	1.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000
	Narrative Elements							
	Tone_speech		Order_event		Narrating_actor		Values_events	
	mean	std	mean	std	mean	std	mean	std
DS	0.792	0.415	0.458	0.509	0.917	0.282	0.750	0.442
Llama	0.583	0.504	0.458	0.509	0.833	0.381	0.583	0.504
Mistral	0.917	0.282	0.958	0.204	1.000	0.000	1.000	0.000
GPT-4o	1.000	0.000	0.958	0.204	0.958	0.204	1.000	0.000
	Narrative Elements							
	Perspective_narrator		Connect_events		Semantic_setting		Temporal_emplotment	
	mean	std	mean	std	mean	std	mean	std
DS	0.833	0.381	0.333	0.482	0.500	0.511	0.375	0.495
Llama	0.333	0.482	0.500	0.511	0.417	0.504	0.375	0.495
Mistral	1.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000
GPT-4o	1.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000

Table 5: Migration Validation Table

the full corpus.

Factual accuracy checks exposed catastrophic errors in some outputs. DeepSeek hallucinated actors, replacing José Manuel Barroso (President of the European Commission) with José Mourinho (a football manager) in 2011 and substituting Jean-Claude Juncker (President of the European Commission) with Joe Biden (President of the USA) in 2015. It also referenced events such as the migration crisis and Brexit before they occurred in the corpus timeline. These errors reflect two distinct failures: the hallucination of unrelated public figures and the insertion of real EU events before they occurred, suggesting the models drew on broader background knowledge of EU politics rather than close engagement with the speeches provided.

At the level of narrative elements, Llama frequently described tropes that were not appropriate to the speech, while Mistral and GPT produced more coherent accounts that connected events and actors through the perspective of the narrating actor. Where no relevant material on the topics appeared in the speeches, some models behaved inappropriately. Mistral, for example, treated the EU "migrating" towards greater integration in 2011 as evidence of a migration narrative. Notably, despite its factual errors in the 300-word analyses, DeepSeek's 600-word migration analysis received the highest quality score in this task, suggesting that the longer narrative was not based on the poor summaries used to create it but rather on its wider knowledge base.

These results demonstrate that systematic human validation, or further computational analysis, is required to detect catastrophic errors. LLMs can support narrative research not by replacing interpretation, but by providing researchers with narrative accounts of political texts at scale. These outputs can be factually grounded by human experts, offering a basis for engaging with large corpora and providing interpretive footholds. A Human-LLM Co-Analysis approach to political narrative research can therefore allow researchers to engage with a broader volume of political text, whilst retaining the human expertise, judgment, and grounding that such research demands. In this sense, LLMs are best understood not as substitutes for interpretive analysis, but as tools that can extend its scale while leaving evaluation and meaning-making in human hands.

	Structural Checks				Factual Accuracy			
LLM	Language		Word_count		Real_events		Key_events	
	mean	std	mean	std	mean	std	mean	std
DS	0.917	0.282	0.250	0.442	0.750	0.442	0.625	0.495
Llama	0.958	0.204	0.375	0.495	1.000	0.000	0.792	0.415
Mistral	1.000	0.000	0.167	0.381	1.000	0.000	1.000	0.000
GPT-4o	1.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000
	Factual Accuracy							
	Real_actors		Key_actors		Real_policy		Key_policy	
	mean	std	mean	std	mean	std	mean	std
DS	0.583	0.504	0.583	0.504	0.583	0.504	0.625	0.495
Llama	1.000	0.000	1.000	0.000	0.958	0.204	1.000	0.000
Mistral	1.000	0.000	1.000	0.000	0.958	0.204	1.000	0.000
GPT-4o	1.000	0.000	1.000	0.000	0.958	0.204	1.000	0.000
	Narrative Elements							
	Tone_speech		Order_event		Narrating_actor		Values_events	
	mean	std	mean	std	mean	std	mean	std
DS	0.375	0.495	0.292	0.464	0.292	0.464	0.333	0.482
Llama	0.500	0.511	0.667	0.482	0.542	0.509	0.542	0.509
Mistral	0.958	0.204	0.958	0.204	0.958	0.204	0.958	0.204
GPT-4o	1.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000
	Narrative Elements							
	Perspective_narrator		Connect_events		Semantic_setting		Temporal_emplotment	
	mean	std	mean	std	mean	std	mean	std
DS	0.125	0.338	0.167	0.381	0.167	0.381	0.167	0.381
Llama	0.500	0.511	0.375	0.495	0.583	0.504	0.375	0.495
Mistral	1.000	0.000	0.958	0.204	1.000	0.000	1.000	0.000
GPT-4o	1.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000

Table 6: Solidarity Validation Table

Model	**Migration Mean**	**Solidarity Mean**
300-Word Analyses		
DS	0.358	0.383
Llama	0.467	0.617
Mistral	0.717	0.817
GPT-4o	0.933	0.883
600-Word Analyses		
DS	0.600	1.000
Llama	0.300	0.100
Mistral	0.600	0.600
GPT-4o	0.600	0.800

Table 7: Narrative Quality Rating Mean Scores: Migration and Solidarity

6. Limitations

It is important to note several limitations of this study. Similarity metrics capture particular dimensions of alignment but cannot, on their own, determine narrative quality. Although the validation stage assesses factual grounding and narrative structure, these evaluations remain interpretive rather than definitive, reinforcing the need to read similarity scores alongside qualitative judgement. The metrics used do not reflect the most recent advances in contextual embedding methods such as BERTScore; however, the divergence between similarity scores and validation outcomes suggests that the case for systematic human validation would likely persist regardless of the specific metrics applied. Specifically for ChatGPT, the 300-word analyses used ChatGPT-4o; following its unavailability, the 600-word analyses used ChatGPT-turbo, introducing a shift in model configuration that limits direct comparability between tasks. Given the rapid evolution of model releases, the results reflect the specific systems examined at the time of analysis rather than fixed capabilities.

The corpus is relatively small (13 speeches), selected to enable detailed examination by human scholars and reflecting a size comparable to interpretive qualitative research, though smaller than typical computational social science datasets. The prominence of EU SOTEU speeches in public discourse means they are likely well represented in model pretraining data, which may inflate similarity scores in ways that are difficult to isolate. The anal-

	Structural Checks				Factual Accuracy			
Code	Language		Word_count		Real_events		Key_events	
	mean	std	mean	std	mean	std	mean	std
DS	1.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000
Llama	1.000	0.000	1.000	0.000	1.000	0.000	0.000	0.000
Mistral	1.000	0.000	1.000	0.000	1.000	0.000	0.500	0.707
GPT-4o	1.000	0.000	1.000	0.000	1.000	0.000	0.500	0.707
	Factual Accuracy							
Code	Real_actors		Key_actors		Real_policy		Key_policy	
	mean	std	mean	std	mean	std	mean	std
DS	1.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000
Llama	1.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000
Mistral	1.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000
GPT-4o	1.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000
	Narrative Elements							
Code	Tone_speech		Order_event		Narrating_actor		Values_events	
	mean	std	mean	std	mean	std	mean	std
DS	1.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000
Llama	0.000	0.000	0.000	0.000	0.500	0.707	1.000	0.000
Mistral	0.500	0.707	1.000	0.000	1.000	0.000	1.000	0.000
GPT-4o	1.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000
	Narrative Elements							
Code	Perspective_narrator		Connect_events		Semantic_setting		Temporal_emplotment	
	mean	std	mean	std	mean	std	mean	std
DS	1.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000
Llama	0.500	0.707	0.500	0.707	0.000	0.000	0.500	0.707
Mistral	1.000	0.000	1.000	0.000	0.500	0.707	1.000	0.000
GPT-4-turbo	1.000	0.000	1.000	0.000	0.500	0.707	1.000	0.000

Table 8: Migration and Solidarity across the full corpus Validation

ysis is limited to English-language texts, and the two narrative scholars share a broadly similar disciplinary background, which may narrow the interpretive range of the human benchmark. While benchmarking against expert scholars provides an appropriate standard for interpretive quality, it should be understood as reflecting one interpretive community rather than the full range of possible analyses. Scholars trained in different traditions or national contexts may produce divergent readings.

These limitations point to productive directions for future research, while also highlighting the value of combining lexical and semantic metrics with structured human validation as a methodological framework for qualitative political narrative analysis using a Human-LLM Co-Analysis approach.

7. Ethics Statement

This study analyses publicly available political texts and does not involve human participants or personal data. Ethical approval was granted by the University of Edinburgh Ethics Board.

8. Acknowledgments

This work was supported by the Volkswagen Stiftung Extraordinary Projects and the EU Horizon 2020 Innovate Projects.

9. Bibliographical References

Muneera Bano, Didar Zowghi, and Jon Whittle. 2023. Exploring qualitative research using llms.

Kenneth Benoit, Scott De Marchi, Conor Laver, Michael Laver, and Jinshuai Ma. 2025. Using large language models to analyze political texts through natural language understanding. *American Journal of Political Science*.

Natalia Chaban, Alister Miskimmon, and Ben O'Loughlin. 2019. Understanding eu crisis diplomacy in the european neighbourhood: strategic narratives and perceptions of the eu in ukraine, israel and palestine. *European Security*, 28(3):235–250.

Thomas Colley and Carolijn {van Noort}. 2022. *Strategic Narratives, Ontological Security and Global Policy: Responses to China's Belt and Road Initiative*. Palgrave Studies in International Relations. Palgrave Macmillan Ltd., United Kingdom.

Niels Dekker, Tobias Kuhn, and Marieke van Erp. 2019. Evaluating named entity recognition tools for extracting social networks from novels. *PeerJ Computer Science*, 5:e189.

Zackary Okun Dunivin. 2025. Scaling hermeneutics: a guide to qualitative coding with llms for reflexive content analysis. *EPJ Data Science*, 14(1):28.

Linus Hagström and Karl Gustafsson. 2019. Narrative power: how storytelling shapes east asian international politics. *Cambridge Review of International Affairs*, 32(4):387–406.

Alexandra Homolar and Oliver Turner. 2024. Narrative alliances: the discursive foundations of international order. *International Affairs*, 100(1):203–220.

Zhichao Hu and Marilyn Walker. 2017. Inferring narrative causality between event pairs in films. In *Proceedings of the 18th annual SIGdial meeting on discourse and dialogue*, pages 342–351.

Muhammad Nihal Hussain, Kiran Kumar Bandeli, Samer Al-khateeb, and Nitin Agarwal. 2018. Analyzing shift in narratives regarding migrants in europe via blogosphere. *Text2Story@ ECIR*, 585:33–40.

Sarah Jenner, Dimitris Raidos, Emma Anderson, Stella Fleetwood, Ben Ainsworth, Kerry Fox, Jana Kreppner, and Mary Barker. 2025. Using large language models for narrative analysis: A novel application of generative ai. *Methods in Psychology*, 12:100183.

Alessandro Maisto. 2025. Collaborative storytelling and llm: A linguistic analysis of automatically-generated role-playing game sessions. *arXiv preprint arXiv:2503.20623*.

Sebastian Michelmann, Manoj Kumar, Kenneth A. Norman, and Mariya Toneva. 2023. Large language models can segment narrative events similarly to humans.

Andrew Piper and Sunyam Bagga. 2025. NarraDetect: An annotated dataset for the task of narrative detection. In *Proceedings of the The 7th Workshop on Narrative Understanding*, pages 1–7, Albuquerque, New Mexico. Association for Computational Linguistics.

Elahe Rahimtoroghi, Jiaqi Wu, Ruimin Wang, Pranav Anand, and Marilyn Walker. 2017. Modelling protagonist goals and desires in first-person narrative. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 360–369.

Georg Rehm, Julian Moreno Schneider, Peter Bourgonje, Ankit Srivastava, Jan Nehring, Armin Berger, Luca König, Sören Räuchle, and Jens Gerth. 2017. Event detection and semantic storytelling: Generating a travelogue from a large collection of personal letters. In *Proceedings of the Events and Stories in the News Workshop*, pages 42–51, Vancouver, Canada. Association for Computational Linguistics.

Paul Ricoeur. 1984. *Time and narrative, Volume 3*, volume 3. University of Chicago press.

Benjamin M. Schmidt. 2015. Plot arceology: A vector-space model of narrative structure. In *2015 IEEE International Conference on Big Data (Big Data)*, pages 1667–1672.

Margaret R. Somers. 1994. The narrative constitution of identity: A relational and network approach. *Theory and Society*, 23(5):605–649.

Evangelia Spiliopoulou, Eduard Hovy, and Teruko Mitamura. 2017. Event detection using frame-semantic parser. In *Proceedings of the Events and Stories in the News Workshop*, pages 15–20, Vancouver, Canada. Association for Computational Linguistics.

Dominik Stammbach, Maria Antoniak, and Elliott Ash. 2022. Heroes, villains, and victims, and GPT-3: Automated extraction of character roles without training data. In *Proceedings of the 4th Workshop of Narrative Understanding (WNU2022)*, pages 47–56, Seattle, United States. Association for Computational Linguistics.

Brent J Steele. 2008. *Ontological security in international relations: Self-identity and the IR state*. Routledge.

Yidan Sun, Qin Chao, and Boyang Li. 2024. Event causality is key to computational story understanding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3493–3511, Mexico City, Mexico. Association for Computational Linguistics.

Nicholas Sloss Treynor and Joshua McCoy. 2025. A case study on user perception of parameterized llm-generated narratives. In *2025 IEEE Conference on Games (CoG)*, pages 1–7. IEEE.

H. White. 1987. *The Content of the Form: Narrative Discourse and Historical Representation*. Johns Hopkins University Press.