ISBN 978-9934-619-52-6

AI in Support of StratCom:

The Use and Evaluation of Large Language Models in Less Widely Used Official EU Languages

PREPARED AND PUBLISHED BY THE

NATO STRATEGIC COMMUNICATIONS CENTRE OF EXCELLENCE

ISBN: 978-9934-619-52-6
Authors: Eduard Barbu, Somnath Banerjee, Tanya Lim, Liene Zīvere
Project Manager: Yukai Zeng
Content Editor: Hadley Newman
Design: Inga Ropša

Cover Image Content generated by AI (GPT-5)

Riga, December 2025

NATO STRATCOM COE
11b Kalnciema iela,
Riga, LV1048, Latvia
stratcomcoe.org
@stratcomcoe

AI in Support of StratCom:

The Use and Evaluation of Large Language Models in Less Widely Used Official EU Languages

Contents

1. Introduction

This report presents a systematic evaluation of contemporary large language models (LLMs) on three critical natural language processing (NLP) tasks: narrative detection, topic modelling, and aspect-based sentiment analysis (ABSA)[1]. These tasks are foundational for understanding and interpreting textual data from diverse sources, including news media, official communications, academic publications, and social platforms. They enable the extraction of meaningful insights that inform public discourse, support communication-related strategic decision-making, and inform communication strategies across domains.

To ensure methodological rigor and domain relevance, the evaluation frame-work was developed in consultation with strategic communications (StratCom) experts. This framework draws on established practices from both NLP research and StratCom analysis, Consequently, the tasks have been adapted from their conventional academic definitions to better reflect the practical and contextual requirements of real-world applications.

The LLMs are evaluated in a zero-shot and few-shot settings[2] on a bilingual dataset comprising documents in both English and Latvian, which enables an assessment of their ability to generalise to domain-specific tasks without additional task-specific fine-tuning. While English benefits from extensive coverage in the training corpora of most models, Latvian poses unique challenges due to its comparatively limited representation in existing datasets[1]. This dual-language evaluation assesses the models' capabilities across both high-resource and low-resource settings, highlighting strengths, limitations, and adaptability in multilingual contexts.

To enhance accessibility, the report is organised in a modular structure in which Section 2 introduces the evaluated LLMs, Sections 4,6 and 7 present the narrative detection, topic modelling, and aspect-based sentiment analysis tasks with evaluation results for both English and Latvian, and Section 8 concludes with a summary of key findings and their broader implications for multilingual NLP and StratCom.

The code, corpora, and prompt templates used in this evaluation are available in a public GitHub repository[2]. This repository supports reproducibility and further research.

1 A fine-grained sentiment analysis approach that identifies specific aspects or features mentioned in text and determines the sentiment expressed toward each each of those aspects separately.

2 Zero-shot evaluation of LLMs test the model's ability to perform tasks without any examples, relying solely on its pre-trained knowledge, while few-shot evaluation uses a small number of examples to guide task performance.

2. Large Language Models

LLMs, trained on large multilingual and multi-domain datasets, can perform tasks such as classification, summarisation, translation, dialogue, and topic extraction, and in this report, we evaluate their performance on narrative detection, topic modelling, and aspect-based sentiment analysis (ASBA) using structured prompt templates that rely on the models' pre-trained knowledge rather than task-specific fine-tuning. Templates are predefined text formats designed to frame instructions in ways that guide the model's reasoning, For example, a sentiment analysis prompt might be framed as: "Classify the sentiment of this review: [text]", while a topic modelling prompt might ask, "Identify the main topic in the following paragraph: [text]"[3].

This prompt-based approach is inherently unsupervised, relying solely on the model's pre-trained knowledge and internal representations rather than on task-specific fine-tuning or annotated training data. It allows for flexible querying of LLMs across different tasks, providing consistent outputs without additional model customisation.

The following LLMs were evaluated in this study:

1. **GPT-4.0**, developed by OpenAI[4], is widely regarded as one of the most advanced transformer-based LLMs available. Trained on a diverse and extensive multilingual corpus, it excels in content generation, reasoning, and contextual analysis. Our evaluation indicates that it performs especially well in high-resource languages such as English and provides competent support for lower-resource languages like Latvian, though results may vary depending on the task and domain, GPT-4.0 was accessed via the Azure OpenAI API[5].

2. **Mistral Nemo** is a variant from Mistral AI[6], optimised for low-latency inference and instruction-following performance. Built on the transformer architecture and trained on high-quality multilingual datasets, it supports various NLP tasks such as summarisation, classification, and structured content generation, Mistral Nemo was accessed via the *OpenRouter API*[7].

3. **Mistral Large** is a high-capacity transformer model from Mistral AI[8] designed for complex reasoning and multilingual understanding, trained on a rich and diverse corpus, and in our evaluation, it demonstrated strong performance in tasks requiring nuanced text generation, abstraction, and structured interpretation. Mistral Large was also accessed through the *OpenRouter API*.

4. **Llama 31-405b**, part of Meta's Llama 3 series[9], is a 405-billionparameter model trained on a wide-ranging multilingual dataset, including books, web data, and academic texts, intended to offer advanced reasoning, high fluency, and contextual coherence. Llama 31-405b was accessed via the *OpenRouter API*.

5. **Gemini Pro**, developed by Google DeepMind[10], is a next-generation multimodal model whose text-only version used in this evaluation builds on the PaLM architecture, is trained on a broad multilingual corpus, and is reported to excel at long-context understanding, reasoning, and few-shot learning, and it was accessed via the OpenRouter API.

All models were accessed through publicly available APIs to ensure consistency, replicability, and ease of integration: GPT-4.0 via Azure OpenAI and the remaining models via the OpenRouter API, which provides unified access to a range of LLMs from various developers.

3. Corpus

3.1. English

For the English-language evaluation, we selected a corpus of 19 articles curated based on their Search Engine Optimisation (SEO) performance and relevance to the target domains, drawn from a range of global and regional outlets to ensure topical diversity and geographic balance including sources such as Euronews 2023[11], BBC[12], AP News[13], Reuters[14], Asahi Shimbun[15], The Straits Times[16], and the South China Morning Post[17].

All articles were either originally written in English or professionally translated, on the assumption that translations preserved narrative structure and key contextual nuances, and while the corpus excludes tabloid sources and highly partisan media to reduce bias, political leanings of sources were not considered, which could introduce skew into the corpus despite efforts to ensure balanced content.

3.1.1. Emerging Themes

A review of the corpus reveals a set of recurring and interrelated themes that are especially relevant for StratCom. These themes emerged from qualitative review/coding of the articles, and provide a framework for evaluating the LLMs' ability to identify and interpret key narrative elements:

1. *Suspected Sabotage* – A prevalent concern across articles, reflecting fears of deliberate attacks on critical infrastructure. This theme is relevant for assessing risks and shaping deterrence messaging.

2. *Investigations and Technical Assessments* – Many articles focus on official investigations, technical findings, and attribution efforts, highlighting the importance of transparency and fact-based communication.

3. *Suspicions and Accusations (Russia, China)* – Several articles raise the possibility of foreign state involvement, underscoring the delicate balance between transparency and diplomatic restraint.

4. *Infrastructure Vulnerability and Security* – Recurring attention is given to the fragility of energy systems, underscoring the importance of resilience and multinational coordination.

5. *Geopolitical Tensions and Diplomatic Reactions* – The broader international context of pipeline incidents and infrastructure damage illustrates the interplay between regional crises and global diplomacy.

6. *NATO Response* – Questions and commentary on NATO's stance and level of preparedness are common, reflecting public expectations for collective action and deterrence.

7. *Energy Security and Financial Implications* – Many articles link physical security concerns to economic risks, pointing to the role of messaging that reassures both the public and the markets.

3.2. Latvian

For the Latvian language evaluation, we selected a corpus of four articles from major Latvian news outlets, including: LSM, Delfi, TVNET, and the news site la.lv. The specific articles were drawn from a range of similar pieces published around the same time, covering the same topic and considering their context. The selected articles included extensive explanations of the event.

All articles were either originally written in Latvian or adapted from English by professional journalists. To reduce bias, the corpus excludes tabloid sources and highly partisan media, focusing instead on mainstream news reporting.

1. *Infrastructure Vulnerability* – Most attention in the Latvian articles focuses on the fragility of energy systems, highlighting critical infrastructure and coordination between different actors.
2. *Investigations and Technical Assessments* – All articles focus on official investigations, and the actors involved, highlighting the role of transparency and fact-based communication.
3. *Energy resources* – Articles mention energy resources as an aspect of security and as a topic of societal importance.
4. *Geopolitical dynamics* – The broader international relations context of pipeline incidents highlights the coordination between different countries and related diplomatic engagement.

3.3. Broader StratCom Considerations

Taken together, these articles address broader strategic concerns such as infrastructure vulnerability, geopolitical instability, and the role of information in shaping public narratives. Beyond the immediate events they describe, they offer insight into how narratives evolve and how they may be exploited by hostile actors.

Malicious actors may seek to manipulate the information environment to disseminate narratives that undermine public trust, sow division, or weaken institutional cohesion. These efforts may be organic or coordinated and can affect strategic communication campaigns, including public perception, electoral dynamics, and international alignment.

By evaluating how effectively LLMs can detect and structure these themes, this report contributes to ongoing efforts to develop tools and methodologies that support resilient and adaptive strategic communication in the context of disinformation and contested narratives.

4. Narrative Detection

4.1. Task Description

The narrative detection task was originally defined in a previous report[18]. This study builds on that foundation incorporating feedback from StratCom experts and insights gained through prior experimentation. Minor refinements have been introduced to improve task instructions and language model performance. This report provides a succinct summary of the updated task; for a comprehensive account of the narrative detection framework, including annotated examples, readers are referred to the earlier report.

In contrast with the original formulation, the revised task now requires both LLMs and human annotators to rank the importance of entities mentioned in the document. Furthermore, the prompt template has been slightly adjusted to elicit more informative and contextually relevant outputs from the models.

Narrative detection[19] refers to the automated analysis of textual data to identify key entities, relationships, and story progression within a narrative[20]. This structured analysis comprises three interrelated layers: **Named Entity Recognition (NER)**, **Relationship Extraction (RE)**, and **Plot Discovery and Story Evolution**. Together, these layers enable a deeper understanding of how narratives unfold over time.

This task evaluates the ability of LLMs to construct these narrative structures. The evaluation is conducted manually by StratCom experts using predefined criteria.

A brief summary of each narrative detection layer is provided here:

1. **Named Entity Recognition (NER)** involves categorising entities in unstructured text into predefined types such as actors, events, locations, and timelines[21]. In our template, we specify the most commonly encountered entity types in narrative detection. Identified entities are also ranked by importance, with scores that reflect their relevance to the narrative.

2. **Relationship Extraction (RE)** builds on the output of the NER layer to identify and classify semantic relationships among entities[22]. These relationships include causal (cause-and-effect links), temporal (chronological sequence of events), spatial (geographic associations), social (interactions and affiliations between actors), and participation (actor involvement in events). This process transforms unstructured text into structured narrative frameworks.

3. **Plot Discovery and Story Evolution** refers to mapping of the overarching narrative arc, often using Freytag's Pyramid: **Exposition**, **Rising Action**, **Climax**, **Falling Action**, and **Resolution**. The interplay of layers begins with **Exposition**, which introduces entities and initial context, triggering the first narrative shift. **Rising Action** follows, as events escalate in complexity through causal relationships. The **Climax** marks a turning point that influences subsequent developments, while the **Falling Action** and **Resolution** describe how entities, relationships and events are resolved. Story evolution complements this structure by identifying pivotal shifts and causal linkages that mark turning points, character changes, or thematic developments.

It is important to note that humans did not originally assign the scores but instead validated model-generated scores. Although this introduced potential bias in entity scoring, the annotators sought to address this by validating the scores against the main article content to assess their alignment with salient narrative elements.

4.2. Template Description

The template defines four narrative detection layers—NER, Relationship Extraction, Plot Discovery, and Story Evolution—along with annotation instructions for both human and machine annotators. It includes few-shot examples for some layers and standardises outputs while retaining the original language of content.

1. NER identifies key entities (people, events, places, timelines), assigns importance scores, and ranks them by relevance.

2. Relationship Extraction captures binary links (e.g., causality, participation) between entities to structure interactions.

3. Plot Discovery maps events using Freytag's Pyramid, aligning narrative stages with previously identified entities.

4. Story Evolution traces major shifts and causal developments in the storyline.

For better LLM performance, Plot Discovery and Story Evolution are treated as distinct layers. Full template details and code are available in the accompanying GitHub repository.

Based on experimental findings, we have split the previously unified final layer into two—Plot Discovery and Story Evolution—to enhance LLM performance in generating structured narrative outputs.

5. Evaluation

5.1. English Results

This section evaluates the performance of five LLMs in narrative detection tasks using English text. The goal is to compare how well these models handle narratives by identifying their strengths, weaknesses, and performance gaps against a human-annotated benchmark.

The evaluation focuses on three key aspects of narrative detection: **Named Entity Recognition (NER)**, **Relationship Extraction (RE)**, and **Plot Discovery and Story Evolution**. These layers cover how the models identify important entities, uncover relationships between them, and track how events develop over time.

Additionally, this section examines at how well the models capture more nuanced geopolitical themes and uncover connections between complex events, which are crucial for building contextually accurate and meaningful narratives.

5.2. Named Entity Recognition (NER)

The NER evaluation focuses on how effectively each model identifies and ranks key entities within a narrative. The models generate importance scores based on entity relevance. Although the initial human annotation did not include numerical importance scores, the annotator cross-referenced the model's output with the main article and qualitatively validated the accuracy of the ranked entities.

MistralLarge MistralLarge was able to capture a large array of entities with relatively good precision. In the Guardian article, for example, it correctly identifies the core actors such as 'Robin Lardot' and 'Balticconnector gas pipeline'. Similarly, in Reuters, it covers important geographical references such as 'Latvia', 'Baltic Sea', and 'Finland', ensuring that the main locations are recognised. However, MistralLarge occasionally introduces extraneous elements that do not enhance the central storyline. In the Guardian article, for example, it marks 'Mediterranean' and 'Helsinki' as significant even though they have minimal relevance to the pipeline investigation. This overinclusion can dilute the otherwise strong performance of the model in identifying critical actors and locales.

Mistral Nemo Mistral Nemo produces relatively consistent and structurally coherent entity lists, by capturing key actors and places with minimal noise. For example, in AP News, the model reliably lists Finnish investigators and key events leading to the anchor's recovery. Despite this strength, Mistral Nemo sometimes omits secondary entities, particularly those related to NATO or alliance-related discussions, even when those entities play a role in shaping the broader geopolitical context. For instance, in the Guardian article, the model accurately flags "Finland" and "Balticconnector" but occasionally misses explicit mentions of NATO or Jens Stoltenberg, restricting its capacity to link these entities to broader strategic themes.

GeminiPro GeminiPro stands out for identifying a wide range of entities, effectively capturing important details in a narrative. In a Guardian article, it correctly picks out key elements like "NewNew Polar Bear," "North Stream pipeline explosions," and "NATO Secretary General Jens Stoltenberg," ensuring that important actors and events are not missed. However, its broad approach often leads to including too many minor details. For instance, in a Reuters article about the Baltic Sea, the model highlights the "Mediterranean

Sea," which is not directly relevant. It also tends to mark smaller mentions, like "Monday's news briefing," as key entities, making it harder to focus on the main story. Another issue is that Gemini Pro sometimes categorises countries like "Finland" simply as locations, missing their changing political roles, such as Finland's recent NATO membership, which is essential to understanding the narrative's context.

Llama 3.1 Llama 3.1 has an impressive breadth in covering entities. In the Guardian investigation into a missing ship's anchor, it flags central entities including 'NewNew Polar Bear,' 'Finnish central criminal police (KRP),' and 'Swedish navy,' thereby successfully charting the various actors involved. However, it frequently becomes verbose, including tangential entities such as 'Mediterranean' or 'tens of thousands of kilometres of undersea infrastructure,' which are mentioned only in passing. This verbosity can obscure the story's main narrative. Llama 3.1 also occasionally categorises significant items too generically, such as labelling "NewNew Polar Bear" simply as a "Chinese vessel". This downplays the context surrounding the investigation. Llama 3.1's tendency to categorise central entities too generically can weaken the contextual clarity of the narrative.

GPT 4.0 GPT 4.0 provides comprehensive entity lists and, in some instances, offers nuanced importance scores that help readers discern focal points of the narrative. In texts like the AP News article on pipeline damage, GPT 4.0 excels at identifying not only primary actors and locations but also associated timelines (e.g., 'reopened on Monday'). However, the model often veers into verbosity, labelling minor phrases such as 'regular news briefing' or 'early Monday' as entities on par with major players like 'Russia' or 'Finnish President Sauli Niinistö.' These extraneous inclusions can overburden the output with details of marginal significance. In some cases, GPT 4.0 also misclassifies major state actors: for instance, it has identified 'Russia' as a location rather than an actor, compromising the geopolitical clarity of the narrative.

Comparison with Human Annotations
Across all these models, human annotations exhibit more targeted precision, focusing on the entities that truly move the narrative forward. Rather than listing every reference encountered in an article, human annotator emphasises actors like 'Finnish central criminal police' or 'Chinese President Xi Jinping' as organisational or geopolitical entities, ensuring clarity around their roles. They also refrain from inflating minor temporal or locational details—such as 'Monday' or 'Mediterranean'—into separate, high-priority entities. This allows the human annotators to be precise in identifying the entities involved, and to provide key context behind why their involvement matters. This helps to highlight the key drivers of the narrative, without including peripheral characters who may not be as important.

Moreover, human annotators are comparatively more accurate in annotating entities in context. LLMs have a tendency to label entities like "Russia" and "Finland" as location instead of actors, which diluted the geopolitical nuances of the narrative. Unlike LLMs, the human annotators were able to sieve out the geopolitical context behind the terms used and label the entities accordingly. This lack of contextual reduced the LLM's accuracy in identifying entities.

5.3. Relationship Extraction (RE)

Next, we look at how well each model identifies and contextualises the relationship between entities. This helps to see how comprehensively the LLM can capture the relationship between entities, thereby enhancing the richness of the geographical context to drive the narrative forward.

MistralLarge MistralLarge was able to map out the direct relationships between entities capturing the investigative and geopolitical relations between them. In The Guardian article, it extracted links such as 'recovers [Finnish investigators, ship's anchor]' and 'is missing [NewNew Polar Bear, front anchor],' which outline the key developments of the investigation. Similarly, in Reuters, it identified 'denies involvement in [Kremlin, Balticconnector pipeline damage]' and 'suggests closing [Edgars Rinkevics, Baltic Sea]'. This allowed it to identify the narrative's most important exchanges on liability and regional security.

Despite this strength, MistralLarge occasionally produced overlapping or repeated relationships that added little nuance. For example, it re-articulated 'investigates [NBI, pipeline damage]' in multiple forms, cluttering the overall analysis. Furthermore, it sometimes stops short of embedding relationships in the broader geopolitical framework. For example, it acknowledged, but did not link NATO's increased patrols to explore the deeper alliance strategies, or to overarching geopolitical themes of energy security.

Mistral Nemo Mistral Nemo tended to yield concise and well-structured relationships. In AP News, it organises the main interactions, such as 'is investigating [Finnish central criminal police, suspected sabotage]', into a logical narrative flow with minimal redundancy. This made it straightforward to identify the cause-and-effect sequences in the article. However, there was sometimes a lack of depth that was needed to capture the geopolitical relationships between entities. For example, in some of the articles, it correctly identified that 'Finland is in NATO', but did not further elaborate on the alliance's operational role or strategic response in patrolling undersea infrastructure. In The Guardian's article, it noted Finland's investigators but omitted explicit references to NATO Secretary General Jens Stoltenberg. This missed out on an opportunity to contextualise Finland's NATO membership within the pipeline sabotage investigation.

Gemini Pro Gemini Pro performs well in detecting direct, surface-level interactions. This includes identifying relationships like 'blaming [Moscow, Nord Stream explosions]' or 'denying [China, sabotage claims].' This improves the visibility of the article's cause-and-effect relationships and ensures that no primary actors or claims go unnoticed.

Nonetheless, there was a tendency to over-include redundant relationships. For example, it reiterated several relationships, such as, "is investigating [Finnish investigators, pipeline damage]" or "contacted [Tallinn, Beijing]" in slightly varied terms. This had the unintended effect of obscuring the key narrative by flooding the output with similar relationships. More notably, Gemini Pro tends to miss the broader geopolitical themes, that provide the key context behind relationships identified. In The Guardian, while it pinpoints "is missing [NewNew Polar Bear, anchor]," it does not always connect these findings to NATO's growing concerns over undersea infrastructure security. This diminishes the ability of the model to highlight larger alliance dynamics or more broadly, the overarching geopolitical themes.

Llama 3.1 Llama 3.1 was able to extract detailed and relevant relationships, that showcased its ability to capture developments within a narrative. In The Guardian, for instance, it outlines "is missing [NewNew Polar Bear, front anchor]" and "recovers [Finnish investigators, ship's anchor]". This added clarity to how the investigation progressed. This

was further complemented by references to external actors, like the “Swedish navy,” which illustrated regional involvement by offering context.

The primary drawback, however, is once again verbosity. Llama 3.1 can restate relationships in multiple forms, offering little to no new information each time. For example, it overproduces variants of “is suspected of sabotage [Chinese container vessel, Baltic connector pipeline]” across multiple lines. Additionally, while it identifies many interactions, it often fails to highlight how these relationships tie into broader themes - such as the significance of Finland’s NATO membership or the alliance’s collective defence posture in the Baltic region.

GPT 4.0 GPT 4.0 achieves broad coverage of relationships, regularly capturing crucial cause-effect links like “is investigating [Finland, pipeline damage]” and “contacted [Tallinn, Beijing].” This approach is useful in building a wider perspective and understanding of who is doing what in the narrative. There is also an interesting nuance of it being able to capture timeline-based relationships, such as “reopened [Baltic connector pipeline, after repairs],” which helps solidify the narrative by piecing the event’s sequence together.

Nevertheless, GPT 4.0 also tends to emphasise repetitive or less impactful ties. For example, it rephrased a pipeline’s reopening multiple times with only slight modifications. Additionally, certain second-tier relationships (which are more nuanced and enrich the geopolitical narrative – like back-channel discussions between NATO officials and Finnish authorities) may remain underexplored. GPT 4.0 pinpoints the “what” quite effectively but can underserve the “why,” particularly regarding strategic alliance dynamics or deeper policy ramifications.

Comparison with Human Annotations
Human annotators provide more contextually rich and geopolitically nuanced relationships. They were able to weave together key themes important to the narrative like alliance politics, strategic infrastructure protection, and energy security. For instance, in *The Guardian*, human annotators succinctly captured “stepped up [NATO, patrols in the Baltic]” and linked it to Finland’s new membership status, spotlighting how a single event can accelerate collective security measures. While most models were able to accurately note an increase in patrols, they often omitted broader themes, like the significance of these actions for NATO’s regional posture or Finland’s integration into the alliance. On the other hand, human annotators were able to distil the key interactions in the narrative and tie them explicitly to broader themes (political, military, and diplomatic implications), thereby linking these relationships to those with strategic value. It is worth noting that this is something that a purely automated approach struggled to achieve.

5.4. Plot Discovery and Story Evolution

In this section, we examine how the model was able to sieve out key events in the narrative and connect them to broader geopolitical themes. Our evaluation assesses how effectively the models can list out chronological developments, and how they can further connect these developments to broader themes.

MistralLarge MistralLarge was able to outline the skeletal structure of the narrative accurately. For instance, in The Guardian, it could chronicle the journey of the pipeline damage, the recovery of the vessel's anchor, and the subsequent suspicion of deliberate sabotage, positioning NATO's increased patrols as a logical falling action. However, its focus on the main incidents often comes at the expense of thematic depth. While it marks the shift toward potential sabotage, it seldom explores its significance for broader alliance politics or energy security. The result is a timeline that is coherent but stops short of integrating the larger strategic narratives that give these incidents broader relevance.

Mistral Nemo Mistral Nemo offers a similarly clear progression of events, which ensures that each stage of the story is concisely presented. In articles like those from AP News, the model punctuates key developments (e.g., the anchor recovery and pipeline damage) with well-structured transitions. Nevertheless, its resolutions tend to be oversimplified. It often concludes with statements like "NATO is committed to protecting undersea critical infrastructure" without examining how this commitment shapes the ongoing discussions on regional security. As a result, this leaves broader questions for readers unexplored – such as whether heightened patrols would signal a doctrinal shift or how they align with Finland's recent NATO membership.

Gemini Pro Gemini Pro effectively delineates a logical flow of events, from the initial discovery of damage through subsequent investigation. For example, in *The Guardian*, it sequences developments in a manner that is straightforward to follow – highlighting milestones like the timeline for pipeline repairs or NATO's immediate reactions. However, Gemini Pro often misses opportunities to tether these incidents to more important themes like energy security, alliance coordination, or the potential implications of sabotage on Europe's broader geopolitical landscape. This omission can make the story feel self-contained, overlooking the strategic reverberations that human annotations frequently underscore.

Llama 3.1 Llama 3.1 often provides a highly detailed event sequence, capturing micro-progressions like the attempts to contact the NewNew Polar Bear or the timeline of pipeline repairs in near-real-time. Yet its verbosity can affect the clarity of the overall story. Enumerating minutiae like secondary references to "tens of thousands of kilometers" of infrastructure tended to bury some of the key themes that drive the narrative's key momentum. Consequently, while the plot arc is traceable, the model's user may need to sift through extensive detail to grasp the heart of the narrative and its thematic resonance.

GPT 4.0 GPT 4.0 provides a balanced overview of each narrative stage, capturing major turning points and offering a coherent sense of progression. In articles like those from Reuters, it efficiently moves from investigative findings to denials of involvement by major actors such as China or Russia. However, GPT 4.0 sometimes leaves questions of geopolitical fallout or shifting alliance dynamics underexplored. Thus, while the story arc is well-defined, the deeper ramifications —for instance, how pipeline sabotage might alter NATO's strategic posture— can remain vague.

Comparison with Human Annotations When compared to the human annotator's work, the models' weaknesses in capturing deeper themes become clearer. The annotator mapped out timelines and included key details, such as repeated attempts to contact

a vessel or coordination between NATO allies, showing how each event connects to the larger geopolitical picture. Rather than treating events as isolated facts, they linked them into ongoing themes, making it easier to understand how short-term events fit into long-term regional strategies — something the models often missed or only partially captured.

5.5. Model-Specific Highlights

The models had distinct strengths and weaknesses which this section aims to explore. Each model was distinctly characteristic, based on their individual performance in various narrative detection tasks. While all models had relative competence in handling English text, their varying approaches to keyword assignment, geopolitical context, and verbosity yielded different trade-offs.

MistralLarge Across the corpus, MistralLarge excelled at filtering out superfluous information, especially in keyword and entity selection. In articles like Reuters and The Guardian, it consistently flags essential actors without cluttering outputs with minor details. However, it occasionally struggles with overlapping topics, especially when events share closely related themes. For instance, references to 'Nord Stream pipeline explosions' and 'Balticconnector pipeline damage' sometimes conflate into repetitive relationship nodes, potentially obscuring each event's distinct narrative importance.

Mistral Nemo Mistral Nemo was distinct in giving well-structured outputs, presenting story elements with clear causal or temporal links. It keeps relationship duplications to a minimum, which aids readability and coherence. Yet the model often underrepresents the broader geopolitical layer of events. For example, while it accurately captures an anchor recovery in *AP News*, it rarely explores how such incidents intersect with Finland's NATO membership or alliance security doctrines. This gap is most evident in stories involving cross-border tensions, where key contextual factors such as NATO's strategic posture remain unexplored.

Gemini Pro Gemini Pro maintains robust clarity in its structured outputs, effectively outlining who does what and when. This clarity extends to naming both primary and secondary actors. However, it tended to underperform in depth and keyword relevance. The model frequently over-included peripheral details, labelling tangential terms or fleeting mentions as if they were central themes. In doing so, it misses the opportunity to highlight more consequential geopolitical elements, like energy security concerns or alliance tensions, that human annotators repeatedly emphasise.

Llama 3.1 Llama 3.1 offered broad coverage, capturing a wide array of actors, events, and locations. In *The Guardian*, for instance, the model enumerates procedural steps and multiple national players in detail. Yet this thoroughness often led to verbosity, which made essential events harder to parse. Llama 3.1 tended to treat less critical mentions in the article with the same level of importance as the core narrative which risks overwhelming readers. Consequently, while the model's scope is impressive, its readability and thematic prioritisation suffer.

GPT 4.0 GPT 4.0 demonstrates high coverage of both entities and relationships, frequently linking cause-and-effect chains to create cohesive narratives. In Reuters articles, it not only identified main events but also noted secondary elements like timelines ("reopened on Monday"), lending a level of granularity that can be beneficial in detailed analyses. However, verbosity and minor themes can detract from the overall focus. Repetitive references to the same event or the inclusion of trivial details (e.g., "regular news briefing") can muddy the key geopolitical currents.

Moreover, while GPT 4.0's "big-picture" potential is evident, it sometimes misses deeper explorations of alliance dynamics or strategic outcomes, falling short of the contextual depth that human annotations consistently provide.

5.6. Latvian Results

This section evaluates the performance of our five LLMs in narrative detection tasks using Latvian text. Similar to the English text discussed earlier, the goal is to compare how well these models handle narratives by identifying their strengths, weaknesses, and performance gaps against a human-annotated benchmark.

5.7. Named Entity Recognition (NER)

GPT 4.0 Overall, GPT 4.0 exhibited strong capabilities in Named Entity Recognition, relationship extraction, and plot discovery, but with notable limitations. The model showed high recognition of key entities with some misclassifications, a fragmented approach to relationship extraction, and a decent but imperfect grasp of narrative structure. It struggled particularly with identifying multiple story shifts and complex geopolitical relationships. Human annotators outperformed GPT 4.0 through their nuanced understanding of context, accurately classifying entities, and identifying implicit story elements. Issues included mislabelling actors like NATO and the EU as locations, and difficulty with relative timelines such as "in the next three months."

Llama 3.1 While Llama 3.1 showed strengths in NER and plot discovery, it struggled with entity classification, relationship extraction, and tracking story evolution. For instance, NATO and the EU were misclassified as locations, and timelines were inconsistently handled. Despite improvements in actor recognition in articles like TVNET, issues remained with overemphasis and simplification. Human annotators demonstrated a better grasp of context, making more nuanced interpretations and identifying complex narrative shifts.

MistralLarge This model reliably identified major entities, particularly locations and actors, across all evaluated articles. However, it misclassified geopolitical organisations and overlooked key figures at times. Temporal expressions like "next April" posed challenges. Its pattern-based recognition worked well for direct mentions but less so for contextual inferences.

Mistral Nemo Mistral Nemo excelled at structured entity recognition like locations but misclassified actors and struggled with complex or relative timelines. Extracted entities often lacked contextual richness, and major participants were occasionally overlooked. Improvements are needed in differentiating entity types and conveying sufficient contextual information.

Gemini Pro The model performs well in identifying most major entities but shows notable limitations. It often misclassifies certain terms, such as interpreting apkures sezona as a timeline rather than a general time frame, and occasionally fails to detect entities altogether. It also struggles to distinguish between actors and locations in contexts involving organisations like NATO and the EU, indicating a need for better contextual understanding. Additionally, the model has difficulty recognising entities implied by the article rather than explicitly stated. Despite these issues, it handles clearly defined entities with reasonable accuracy.

Comparison with Human Annotations Human annotators demonstrated superior contextual awareness, correctly identifying implicit and complex entities that the models overlooked.

5.8. Relationship Extraction (RE)

GPT 4.0 In Latvian, GPT 4.0 displayed limited depth in relationship extraction. It occasionally merged entities whose roles evolved separately and often stayed confined to one section of text. Relationships were identified but not elaborated or contextualised, sometimes emphasising irrelevant links. Human annotators, on the other hand, understood narrative structures and pinpointed crucial entity relationships that shaped the story.

Llama 3.1 Llama 3.1 generally captured entity links but lacked interpretive depth. Surface-level connections dominated, and more complex links were missed or wrongly inferred. Human annotators offered more coherent and grounded relationship maps.

MistralLarge This model identified fundamental relationships but lacked precision. Generated connections were often vague or assumed without textual support. Misinterpreted relationships limited narrative coherence.

Mistral Nemo Mistral Nemo found some valid relationships but missed many important ones. Even identified links were poorly contextualised. Human annotators succeeded by interpreting the context and assigning interaction types (e.g., alliance, conflict).

Gemini Pro The relationships extracted by the model are generally simple and easy to understand, but they lack depth and nuance. The model does not engage in a deeper analysis of the article to uncover more complex connections between entities. It shows limited textual understanding, often recognising only a few basic and explicitly stated connections. For instance, in one Latvian article, it identified only four possible relationship keywords, all of which reflected very straightforward associations.

Comparison with Human Annotations Human annotations captured more complete and nuanced relationships, distinguishing between evolving dynamics and static connections. This depth of interpretation greatly improved overall narrative understanding.

5.9. Plot Discovery

GPT 4.0 GPT 4.0 captured central plotlines reasonably well across articles. It aligned with human interpretations in some (e.g., LA article), but in others, like Delfi, it often missed the story's climax or over-emphasising the initial incident. It simplified plot progression, lacked depth, and struggled o detecting subplots and thematic layers.

Llama 3.1 Llama 3.1 consistently delivered coherent but over-condensed plot structures. It often truncated after the climax, overlooking developments in the resolution phase. NATO and EU discussions dominated its focus at the cost of wider thematic elements.

MistralLarge Plot coverage was solid but thematically skewed, prioritising specific entities over the natural flow of events. Redundancy and inconsistent sequencing weakened overall plot fidelity. The model would benefit from stronger mechanisms for reconstructing article progression in a manner that reflects the full narrative arc.

Mistral Nemo Key events were identified, but plot sequencing remained weak. Important developments were underemphasised, while less relevant events were given undue attention. Logical flow and event significance were not well prioritised.

Gemini Pro The model often focused on a narrow slice of the article, leading to mismatches in how information is assigned across sections. This constrained focus resulted in

an incomplete understanding, as significant portions of the article are insufficiently considered. It frequently drew conclusions from roughly one-third of the text, failing to account for the article's overall structure and message.

Comparison with Human Annotations Human annotations demonstrated deeper thematic understanding, clearer narrative arcs, and correctly sequenced plot events. They integrated subplots and transitions that models tend to oversimplify or omit.

5.10. Story Evolution

GPT 4.0 Recognised only one shift per article, often treating exposition as a shift. Broader narrative evolution was overlooked, especially in political and energy-related articles. Human annotators perceived multiple transitions and situated them within geopolitical frameworks.

Llama 3.1 Identified main shift but struggled with secondary developments. For example, in *LSM*, the model caught only the final shift. Its recognition of geopolitical importance was present but remained shallow.

MistralLarge Often mistook introductory sections as turning points. Missed causal links and underrepresented evolving relationships. Some shifts were correctly identified but lacked a holistic view.

Mistral Nemo Story progression was poorly tracked. Changes in narrative were either missed or oversimplified. Central themes were wrongly framed as story shifts, distorting the broader structure.

Gemini Pro The model successfully detected a topic shift, and its identification aligns well with the structure of the article. However, it focuses only on the initial section, failing to recognise any additional shifts that may occur later in the text. Causal relationship extraction was handled well, with at least one relevant aspect correctly identified. The model identifies more than a single causal link, suggesting relatively better performance in this area. Nevertheless, the extracted relationships remain simplistic, though not incorrect.

Comparison with Human Annotations Human annotators captured the development and transformation of stories more completely. They tracked causal progressions and relational dynamics, offering a fuller understanding of how narratives unfold over time.

5.11. Conclusions on Narrative Detection Task

The narrative detection task demonstrated that while LLMs can effectively extract entities, relationships, and plot structures in English-language texts, their performance in Latvian remains significantly more limited. Across all evaluated layers—Named Entity Recognition (NER), Relationship Extraction (RE), Plot Discovery, and Story Evolution—English outputs showed higher fluency, stronger contextual integration, and greater alignment with human annotations.

Models like **GPT 4.0** and **MistralLarge** performed consistently well in English, accurately identifying entities and constructing coherent narrative arcs. They showed strength in relationship extraction and timeline development, albeit with issues of verbosity and occasional redundancy. In contrast, their Latvian counterparts struggled with entity classification (especially distinguishing actors from locations), timeline handling, and thematic depth. Errors such as mislabelling geopolitical organisations, under-representing story shifts, and neglecting subplots were common in Latvian outputs.

Despite these challenges, Latvian model outputs were not without merit. **Llama 3.1** and **GPT 4.0** in particular showed promise in capturing basic plot sequences and detecting some causal relationships, although their depth of interpretation and narrative cohesion lagged behind English performance. Human annotators consistently outperformed all models, particularly in identifying implicit relationships and aligning story elements with broader geopolitical frameworks.

The bilingual comparison highlights a persistent gap in model performance between high-resource and low-resource languages. This reinforces the need for domain adaptation, template tuning, and potentially additional training data to enhance LLM performance in languages like Latvian. At the same time, it underlines the value of human oversight in complex narrative analysis tasks, especially when strategic interpretation and geopolitical nuance are critical.

6. Topic Modelling

Topic modelling[23] is an unsupervised machine learning technique used to identify and extract the underlying themes or topics from large collections of text data. It automatically organises and summarises documents by finding patterns of word co-occurrence. Newer approaches to topic modelling[24] harness the power of LLMs, such as BERT[25] and GPT[26]. By using contextual embeddings and transformer-based methods, they produce meaningful topic clusters. Topic Modelling, as defined in the literature, serves as the foundation for the task-oriented definition presented in the next section.

6.1. Task Definition

Traditional topic modelling involves computing a probability distribution over words in the vocabulary to identify key topics within a large corpus. We have adapted this approach to align with the annotation practices employed by the NATO StratCom COE, focusing only on the most significant words within a topic.

Rather than using expansive corpora typical of traditional topic modelling algorithms, we leverage LLMs pre-trained on vast collections of documents. Our corpus consists of articles from various news outlets, handpicked by a NATO StratCom COE expert for their relevance and analytical value.

Both the NATO StratCom COE expert and the selected LLM annotate each article using a structured template designed to ensure consistent and systematic annotation. This template provides clear criteria for evaluating and assigning topics. The goal is to label each article with topics from a predefined set, similar to a supervised topic modelling approach. To challenge the annotators, we include topics that are closely aligned with the article's central themes as well as topics that are unrelated. Once the LLM or human annotator assigns the appropriate topics, they must select the most relevant keywords for each significant topic.

6.2. Template Description for Topic Annotation

Both the human annotator and the LLM use the following template instructions for assigning topics and keywords.

1. **Predefined topics.** The expert assigns each article a set of predefined topics, for example with 7 topics designated for the corpus. We generate a set of 7 topics similar to the expert's and 7 deliberately dissimilar, with the goal of testing whether the LLMs can accurately differentiate legitimate topics from fabricated ones. Additionally, each relevant topic is assigned a weight from 1 to 5:
    - 1: Minimally relevant
    - 5: Highly relevant
2. **Keywords**. Unlike traditional topic modelling, where keywords are individual words or multi-word phrases, the keywords we assign are free expressions that capture significant concepts within the text. For each assigned topic, the most relevant n-grams (keyword phrases) are extracted from the article and weighted based on their relevance and frequency. A score of 3 is the maximum the model can receive.

- Score 3: Highly relevant, central keywords that are frequently mentioned.
- Score 2: Relevant keywords that are important but mentioned less frequently or are less central.
- Score 1: Keywords that are loosely related or briefly mentioned.

3. **Total topic score**. The weights of all keywords are added up to calculate a total topic score, which represents the significance of the topic within the article.
4. **Normalised significance for comparative analysis.** To enable straightforward comparison between topics, the total topic score is normalised on a scale from 0 to 10. This normalisation is achieved using the formula (1):

$$\left(\frac{\text{Total Weight}}{\text{Maximum Possible Weight}} \right) \times 10 \qquad (1)$$

To illustrate this methodology, consider the topic **"Suspected Sabotage"** annotated with high relevance with the normalised score specified in the template above.

- Topic: Suspected Sabotage
- Weight: 5
- Keywords:

– "damage to a gas pipeline" = 3
– "they were looking into the Chinese vessel, the New Polar Bear, and a Russian-flagged ship, the Sevmorput" = 3
– "as well as other vessels present" = 3
– "incident was due to "outside activity"" = 3
– "A Norwegian Navy ship shadowed a Chinese container ship investigated over damage to a gas pipeline" = 2
– "KV Sortland followed the New Polar Bear" = 2

- Total Weight: 16
- Normalised significance: 8.0

6.3. Evaluation

We conduct two types of evaluation:

1. Quantitative evaluation, where we automatically compare the topics and keywords identified by the LLM with those provided by a human expert to measure overlap and alignment.
2. Qualitative evaluation, where experts manually review selected articles and assess the LLM's outputs in depth, focusing on accuracy, relevance, and insight.

As before, we conduct the evaluation for English and Latvian. The methodology for both languages is identical.

6.3.1. English

Table 1 presents a comparative evaluation of five LLMs against human annotations of our English corpus. The evaluation is based on two criteria:

Model	CT	HT	MT	CK	HK	MK
LLaMA 3.1	78	96	94	19	354	388
Mistral Nemo	67	96	79	7	307	272
Mistral Large	76	96	115	24	349	389
GPT 4.0	66	96	77	24	310	277
GeminiPro	66	96	89	2	306	535

TABLE 1. Comparison of LLM outputs with human annotations on 14 English news articles. CT = Common Topics, HT = Topics assigned by Humans, MT = Topics assigned by Machines, CK = Common Keywords, HK = Keywords by Humans, MK = Keywords by Machines.

Topic-Level Comparison:

- CT (Common Topics) – the number of overlapping topics identified by both the LLM and the human expert.
- HT (Human Topics) – the total number of topics identified by the human.
- MT (Machine Topics) – the total number of topics identified by the LLM.

Keyword-Level Comparison:

- CK (Common Keywords) – keywords identified by both the human and the LLM.
- HK (Human Keywords) – the number of keywords selected by the human annotator.
- MK (Machine Keywords) – the number of keywords generated by the LLM.

Llama 3.1 and Mistral Large show the highest alignment with human topic annotations. Mistral Large and GPT 4.0 perform best in keyword overlap. Interestingly, Gemini Pro assigns the largest number of keywords (535), but only 2 match those identified by humans, indicating overgeneration or misalignment. These results highlight the need for balance between coverage and interpretive accuracy in LLM outputs.

Although the quantitative evaluation provides a general sense of model performance, it cannot assess whether the topics selected by an LLM—but not identified by the human expert—might still be meaningful or relevant. Similarly, the keyword comparison focuses on exact matches, yet partial or semantic overlaps between machine- and human-selected keywords may still reflect valuable alignment. To address these limitations, our expert selected one document from the corpus for a more in-depth qualitative evaluation.

6.3.2. Model-by-Model Analysis: Al Jazeera Article

To deepen our understanding of the models' performance, we conducted a qualitative comparison using a representative article from Al Jazeera. The article addressed the Balticconnector pipeline incident and its implications for key themes such as regional security and national infrastructure. Each model was evaluated on its ability to: (i) identify and prioritise central themes, (ii) select relevant and semantically meaningful keywords, (iii) structure outputs clearly, and (iv) align with both core and peripheral narratives.

Mistral Nemo Mistral Nemo produced a clean, logically ordered output that prioritised central themes such as *Suspected Sabotage* and *Energy Security,* using relevant phrases like "suspected sabotage of subsea gas pipeline" and "heavy object was found near the pipeline damage." The structure was readable, and topic weights were clearly defined, making its analysis easy to interpret. It also performed well on *Infrastructure Vulnerability*, correctly capturing terms like "limited number of energy links to the rest of the bloc" and "gas pipeline and telecoms cable were broken." However, the model offered limited elaboration on secondary themes such as *NATO Response* and *International Cooperation*, missing nuances like Tallinn's diplomatic outreach to Beijing or the Latvian president's remarks about closing the Baltic Sea. Overall, Mistral Nemo excels at structured topic detection but could benefit from broader geopolitical contextualisation.

Mistral Large Mistral Large was accurate in defining topics, identifying themes like *Suspected Sabotage*, and identifying supporting evidence like "external mechanical force" and "investigation into Chinese ship." It avoided redundancy in keyword selection and effectively distinguished between closely related topics such as *Energy Security* and *Infrastructure Vulnerability*. This helped prevent conceptual overlap and maintained high interpretability. That said, some secondary topics like *NATO Response* and *International Cooperation* received comparatively less attention, potentially missing connections such as "Tallinn has contacted Beijing" or China's call for a "professional investigation." The model is well suited for use cases where clarity and precision are paramount, though its output could be strengthened by more explicit engagement with diplomatic and strategic subtexts.

Llama 31-405b Llama 31-405b produced a broad and detailed topic map, strongly identifying *Suspected Sabotage*, *Energy Security*, and *Infrastructure Vulnerability*. Its keyword choices — such as "investigation into suspected sabotage and "limited number of energy links" — provided useful specificity, while its inclusion of topics like *International Cooperation* captured important diplomatic references, such as "China and Finland have begun communication" and "jointly safeguard cross-border infrastructure." However,

verbosity occasionally undermined clarity; phrases like “they were looking into two ships concerning the incident” added bulk without enhancing semantic precision. Additionally, the model’s high weighting of infrastructure-related topics may have slightly skewed the overall thematic balance. Llama was effective in terms of identifying key relevant evidence supporting the topics but sometimes became a bit excessive in its elaboration. Human annotation should complement the model’s performance to ensure a comprehensive and precise reconstruction of its narrative.

GPT 4.0 GPT 4.0 demonstrated high sensitivity to narrative detail and broad thematic coverage, especially in extracting keywords for topics such as *Suspected Sabotage*, where it included unique entries like “NewNew Polar Bear” and “telecom cables cut.” It effectively highlighted secondary topics such as *Environmental and Technical Challenges*, referencing poor sea conditions and inconclusive investigations. These were details that the other models sometimes omitted. It placed emphasis on keywords like “external mechanical force” and “confirmed damage”, which created a strong semantic alignment with the article’s investigative tone. However, the model was prone to verbosity and occasional repetition, a pattern observed in several models. This tendency sometimes overshadowed more salient diplomatic developments, such as “Tallinn contacted Beijing” or “China denies role”. GPT 4.0 was particularly useful in identifying an in-depth, fine-grained keyword extraction, but it requires streamlining and vetting by human annotators to improve its effectiveness.

Gemini Pro Gemini Pro produced well-structured and highly readable outputs, correctly prioritising *Suspected Sabotage* and *Energy Security* with straightforward keywords such as “incident stoked concern about the security of energy supplies.” It also surfaced interesting secondary themes like *Environmental and Technical Challenges*, which helped to offer a broader perspective on investigative difficulties and infrastructure conditions. However, it introduced less relevant topics such as *Covert Disruptions*, which detracted from the core storyline and introduced speculative framing. Additionally, Gemini Pro failed to capture key actors and phrases—such as “NewNew Polar Bear” or “Kremlin’s response”—that were essential to the incident’s geopolitical context. Hence, although Gemini Pro performed well as a tool for summarising the article’s main content and to cleanly frame the topics involved, it overlooked nuances in the article, which makes it less relevant and effective for in-depth analysis where identifying high-impact terminology is critical.

Summary Overall, the models exhibited distinct strengths. Mistral Large and Llama appear more suitable for structured thematic exploration, where the themes are identified by a human annotator beforehand, while GPT-4.0 is better suited to combing through the article to extract key details. It is worth noting that in terms of reader readability, Mistral Nemo performed well, offering precise structured outputs, while Gemini Pro’s clarity also stood out. However, all models’ efficacy would be enhanced by human oversight and a clear analytical objective. Models should be chosen based on the operator’s specific analytical goals.

6.3.3. Latvian

Table 2 shows a quantitative comparison of five LLMs on Latvian articles, using the same evaluation metrics as in the English analysis.

Model	CT	HT	MT	CK	HK	MK
LLaMA 3.1	8	22	20	0	26	31
Mistral Nemo	5	22	17	0	18	21
Mistral Large	15	22	33	4	46	68
GPT 4.0	12	22	26	2	37	51
GeminiPro	13	22	38	0	40	53

TABLE 2. Comparison of LLM outputs with human annotations on 4 Latvian news articles.
CT = Common Topics, HT = Topics assigned by Humans, MT = Topics assigned by Machines,
CK = Common Keywords, HK = Keywords by Humans, MK = Keywords by Machines.

Topic-Level Comparison:

- CT (Common Topics) – number of topics identified by both the human
- expert and the LLM.
- HT (Human Topics) – total number of topics annotated by human experts.
- MT (Machine Topics) – total number of topics generated by the LLM.

Keyword-Level Comparison:

- CK (Common Keywords) – overlapping keywords between human annotations
- and LLM outputs.
- HK (Human Keywords) – number of keywords assigned by humans.
- MK (Machine Keywords) – number of keywords generated by LLMs.

To deepen our understanding of the models' performance, we conducted a qualitative comparison using four selected Latvian articles from Delfi, TVNET, LSM and LA.LV. The articles addressed the Balticconnector pipeline incident and its implications for key themes such as regional security and national infrastructure. Each model was evaluated on its ability to: (i) identify and prioritise central themes, (ii) select relevant and semantically meaningful keywords, (iii) structure outputs clearly, and (iv) align with both core and peripheral narratives. As similar themes appeared and model performance for all articles were quite similar, the qualitative evaluation was therefore conducted across all four articles, to capture recurring patterns.

Mistral Large stands out as the top performer in both topic and keyword overlap, identifying 15 topics and 4 keywords in common with the human annotators.

GPT 4.0 also shows relatively strong alignment, with twelve shared topics and modest keyword overlap. Llama 3.1 and Gemini Pro demonstrate decent topic identification capabilities but show zero keyword alignment, which indicates either divergence in lexical expression or semantic representation. Mistral Nemo, while more conservative in its outputs, had the lowest alignment with human topics and no shared keywords.

These results suggest that while the models are capable of identifying major themes in Latvian text, the precision of keyword matching remains limited. As with English, a

qualitative evaluation is required to determine whether the models' divergent outputs contain thematically valid insights not captured in the human annotations.

Mistral Nemo While other model sometimes selected topics that differed from those chosen by human annotators, it still demonstrated a clear thematic understanding of the articles. Mistral Nemo's topic selections are mostly appropriate and generally well aligned with the content. However, two specific topic choices raise concerns: "Cultural Heritage Preservation" in the TVNET article and "Criminal Inquiry" in another evaluation appear contextually inaccurate, suggesting the model did not fully grasp the article's intent. Despite this, topic weightings are appropriate and reflect the importance of each theme. A notable inconsistency occurs in the LA article, where the model merged two topics, omitted keyword weights, yet proceeded to calculate normalised significance—raising questions about its internal logic. Additionally, while the keywords are not repetitive, they are overly shortened; including more contextual terms would enhance interpretability and clarify their relevance to each topic. As with previous models, this one also demonstrates significant weaknesses in applying mathematical formulas accurately.

Mistral Large Mistral large performed appropriately in identifying the correct topics and prioritising them in a manner that generally reflects the content of each article. All selected topics are relevant; however, some could be merged due to the similarity of their associated keywords. Compared to other models, this one demonstrates a closer alignment with human evaluative practices. For instance, in the LA article, the keywords associated with the topic "Natural Gas Infrastructure" closely match those identified by the human annotator, indicating a similar understanding of the text. While some keywords differ slightly from those chosen by the human, they remain contextually appropriate and do not detract from the overall coherence. The keyword weightings are also largely in line with human assessments. Nevertheless, the model exhibits significant weaknesses in its handling of mathematical components. While basic counting appears accurate, the more complex numerical aspects, such as the use of formulas, are poorly executed.

Llama 31-405b Llama produced strong results by selecting relevant topics for each article, such as Infrastructure Vulnerability, and assigning weights that generally reflected the content accurately. A notable pattern observed across three articles is the repeated inclusion of the specific topic "Economic Ties in Northern Europe." While this is not an incorrect classification, it is unusually specific and unlikely to be chosen by a human annotator as one of the few primary topics. The model also performed well in keyword extraction: although human annotators typically considered the full context of the article, the model successfully narrowed down keywords in a manner that was contextually appropriate. For example, the keywords related to Energy Resources were correctly aligned with the topic and weighted similarly to human assessments. Overall, the model's textual analysis demonstrated good alignment with the source material. However, significant issues were identified in the mathematical component; specifically, the calculation of normalised significance was incorrect in nearly all cases across the four articles, with only one topic being calculated correctly.

GPT 4.0 ChatGPT 4.0 Across all evaluated articles (Delfi, LA, LSM, and TVNET), demonstrated a generally sound ability to identify thematically relevant topics and keywords, with reasonable alignment to the article content. However, it frequently exhibited issues with keyword categorisation, overgeneration of topics, and unclear or inconsistent topic weighting—particularly in relation to complex geopolitical and infrastructural concepts. Although many of the divergences from human annotations were semantically valid, the model's logic for assigning weights and applying quantitative formulas remained opaque or flawed. Overall, while the model shows potential for supporting fine-grained topic analysis, it requires refinement in

conceptual classification, weighting methodology, and mathematical reasoning to achieve higher reliability in academic or analytical contexts.

Gemini Pro Gemini Pro across all four articles (Delfi, LA, LSM, and TVNET), showed a generally good ability to identify relevant topics and keywords, but often generated too many topics—up to 10 in cases like LA and TVNET—where fewer would have sufficed. Topic overlap and fragmentation were common; for instance, in Delfi, “international cooperation” and “economic ties in Northern Europe” were treated as separate topics despite sharing nearly identical keywords. While some keywords were accurate, many were overly short or repeated, such as in TVNET where the same word appeared three times under one topic, reducing clarity and distinction. Weighting was also inconsistent— “energy resources” in Delfi, for example, was weighted more heavily by the model than by the human, suggesting a more surface-level interpretation. In several cases, topic choices like “disaster impact” in Delfi were considered contextually inappropriate, as the article did not describe a disaster scenario. Across all articles, the model struggled with applying mathematical formulas, with incorrect normalised significance scores and unexplained figures like “20” noted in Delfi. Despite these issues, the selected topics and keywords were relevant and aligned partially with human interpretations. Further refinement is needed in topic consolidation, keyword contextualisation, and formulaic logic to enhance reliability and analytical quality.

Summary The models effectively identify relevant topics and generally align with human assessments, particularly in areas like Infrastructure Vulnerability and Energy Resources. However, they often generates too many topics, fragmenting related themes and sometimes misunderstanding broader contexts, as seen with topics like ”Economic Ties in Northern Europe” and ”Cultural Heritage Preservation.” Keywords are accurate but overly shortened, reducing clarity, and some are repetitive within topics. While topic and keyword weightings are mostly appropriate, inconsistencies exist, and the model struggles with mathematical formulas, particularly in calculating normalised significance. Despite these issues, the model provides useful insights, with room for improvement in keyword clarity, topic merging, and formula accuracy.

6.4. Conclusion

The comparative evaluation of LLMs across English and Latvian datasets reveals both capabilities and challenges in applying topic modelling for multilingual analysis. While all five models demonstrated some level of alignment with human annotations, their performance varied significantly depending on the language and the evaluation metric.

For English, Llama 3.1 and Mistral Large exhibited the strongest topic-level alignment, with high counts of shared topics and relatively balanced topic generation. Mistral Large and GPT 4.0 stood out in keyword overlap, suggesting an ability to capture fine-grained semantic detail. Gemini Pro, despite producing the largest volume of keywords, showed minimal overlap with human annotations, signaling overgeneration and possible misalignment with expert interpretations.

In contrast, the Latvian results showed generally lower overlap, especially in keyword matching, where most models struggled to replicate human-selected expressions. Mistral Large again performed best in both topic and keyword metrics, indicating its robustness across languages. GPT 4.0 followed closely but encountered challenges with weight assignment and formulaic consistency. Llama 3.1 and Gemini Pro showed acceptable topic identification but no meaningful keyword alignment. Mistral Nemo displayed the weakest alignment overall, particularly in

Latvian, suggesting that a more conservative topic generation strategy may not transfer well across languages.

Qualitative assessments reinforce the quantitative findings. English evaluations showed that models could capture central themes, yet often struggled with nuance, narrative framing, and over-repetition. Latvian evaluations highlighted similar issues, compounded by frequent inaccuracies in mathematical calculations (e.g., normalised significance scores), fragmented topic structures, and overly brief or repetitive keyword choices.

While unsupervised LLMs show considerable potential for topic modelling in both English and Latvian, their outputs are most effective when guided by human oversight. Future work should focus on improving cross-linguistic consistency, refining formulaic reasoning, and enhancing contextual sensitivity, especially in underrepresented languages.

7. Sentiment Analysis

7.1. Task Description

Sentiment analysis is a core task in natural language processing (NLP) that aims to determine the polarity—positive, negative, or neutral—expressed within a text. While traditional sentiment analysis assigns sentiment at the document or sentence level, Aspect-Based Sentiment Analysis (ABSA) offers a finer-grained alternative. ABSA focuses on identifying sentiments tied to specific aspects of an entity, providing more targeted insight into user opinions[27,28,29].

Unlike standard methods, ABSA captures nuanced and sometimes implicit expressions of sentiment, which can be influenced by linguistic complexity, contextual factors, or negation[30]. A key challenge in ABSA lies in detecting these implicit sentiments, where aspect terms are not overtly mentioned but must be inferred from context.

This is especially relevant for analysing media narratives in the context of information warfare, where adversarial actors attempt to shape public opinion. ABSA can support intelligence and policy stakeholders in identifying sentiment trends, uncovering disinformation patterns, and reinforcing strategic communications. In multilingual environments, ABSA enables detection of cross-lingual disinformation campaigns by identifying sentiment variations across languages[31,32]. When integrated into broader cybersecurity frameworks, ABSA helps monitor real-time narrative shifts and deploy counter-narratives that preserve alliance cohesion[33].

Given its utility for both operational awareness and public trust, ABSA emerges as a powerful analytical tool. In this study, we focus specifically on implicit ABSA, targeting subtleties that conventional methods may overlook. In the scope of ABSA, the following areas have been studied:

- Proposing implicit aspects of predefined Named Entity (NE) categories (section 5.7 above) using a semi-automatic approach.
- Leveraging LLMs to extract these implicit aspects
- Evaluating LLMs' performance in ABSA across two different language settings:
  (i) High-resourced language (English) and
  (ii) a low-resourced language (Latvian)

7.2. Defining Implicit Aspects

Identifying implicit aspects is challenging due to their contextual ambiguity, lack of explicit cues, and varied semantic interpretation. To address this, we developed a semi-automatic method for aspect identification.

The process began by prompting four LLMs to generate candidate implicit aspects for each entity category. These LLMs – chosen for their architectural diversity – provided broad coverage of possible interpretations.

We then used OpenAI's GPT-4 to analyse and consolidate these outputs. GPT-4 helped unify overlapping aspects, resolve inconsistencies, and improve the quality of aspect proposals.

Finally, two domain experts manually reviewed and refined the aspects. Their role was to validate semantic relevance and contextual coherence, ensuring alignment with the study's analytical goals.

This semi-automatic pipeline combines LLM generalisation with human domain expertise, resulting in more robust and reliable aspect identification. It allows for a richer capture of implicit semantic relationships, which are essential for more in-depth sentiment analysis.

Following the semi-automatic approach, we identified 13, 11, and 15 aspects for location, actor, and event entities, respectively.

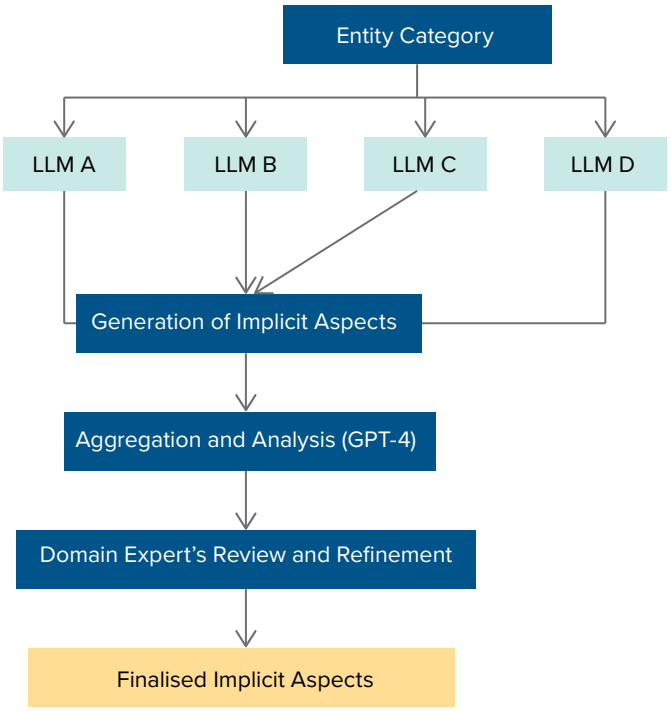

FIGURE 1. Semi-Automatic Pipeline for Identifying Implicit Aspects

7.3. Template Description for ABSA

Given an article with identified named entities (NEs) and a predefined set of implicit aspects, we designed a prompt template to guide the LLM through the sentiment analysis process.

Recognising the inherent ambiguity in identifying implicit aspects, our methodology employed a structured prompt-based approach that leverages the Chain-of-Thought (CoT) reasoning technique. CoT enables the model to engage in step-by-step reasoning, improving its ability to interpret more nuanced, context-dependent sentiment expressions. Decomposing the task into smaller, interpretable steps allowed the CoT framework to enhance the model's precision when evaluating sentiments tied to specific implicit aspects. This ensures greater transparency in the process and consistency in judging sentiments across varying contexts.

For each entity under one of the three categories —*Location*, *Event*, or *Actor*—, the model is guided through the following structured steps:

1. Identify the relevant implicit aspect terms from the predefined set (see Section 7.2).

2. Extract textual excerpts from the article that suggest the presence of the aspect.

3. Justify how the excerpt implies the identified aspect.

4. Assign a Confidence Level (High, Medium, Low) to reflect certainty in the inference. This helps assess the model's confidence in its output.

5. Determine the Polarity of the aspect (Positive, Negative, Neutral).

6. Quantify sentiment strength using a Polarity Score on a 0.1 - 1.0 scale.

This prompt design ensures that the LLM not only identifies and evaluates sentiment, but also explains its reasoning in a structured, interpretable manner that can be cross validated by domain experts.

Figure 2 illustrates a sample output generated by one of the LLMs using this prompt on a selected article from our corpus:

Category: ACTOR

Entity: Petteri Orpo
Aspect: Communication

- **Excerpt:** *Petteri Orpo added that the cause was not yet clear.*
- **Explanation:** Petteri Orpo communicates about the uncertain cause of the damage, demonstrating his role in conveying information.
- **Confidence Level:** High
- **Polarity:** Neutral
- **Polarity Score:** 0.5

EXPERT ANNOTATION: *CORRECT*

FIGURE 2. Sample output for an entity in a English article

7.4. Evaluation

We assessed the performance of five LLMs —GPT, Mistral Large, Mistral-nemo, Llama-3.1, and Gemini Pro— on the ABSA task in zero-shot settings, where no examples were provided to the LLMs in advance. The evaluation was conducted using both quantitative and qualitative methods.

Due to the inherently implicit and interpretive nature of aspect-based sentiment analysis (ABSA), a fixed gold standard was not feasible. Instead, the authors manually reviewed the outputs to assess each (Entity, Aspect, Polarity) triplet generated by the models. These were evaluated across five different

criteria: the relevance of the entity-aspect pairing, the coherence of the explanatory rationale, the appropriateness of the assigned sentiment, and the internal consistency of the triplet when considered as a whole.

Each output followed a common format, consisting of the identified entity, its category (Actor, Location, or Event), an implicit aspect from a predefined set, a representative excerpt, a free-text explanation, and a polarity assignment accompanied by a confidence score. Annotators judged the triplet as **CORRECT** only if all components were logically sound, mutually coherent, and contextually plausible. If any part of the explanation, sentiment, or aspect linkage was flawed—even marginally—the triplet was marked **INCORRECT**.

This approach emphasised interpretive validity over superficial accuracy. Excerpts were read in light of their broader article context, and judgments were grounded in whether the explanation plausibly supported the aspect and sentiment claimed. This required annotators to balance surface-level verification with deeper interpretive reasoning, moving between verifying surface-level matches (e.g., the presence of the named entity) and deeper narrative reasoning (e.g., does the excerpt implicitly reflect international tension, trust, or uncertainty?).

This ensured a high degree of comparability across models, while allowing for the annotator's discretion in judging whether the model was valid or over-speculating.

7.4.1. English

For English, two authors manually reviewed the outputs to assess each (Entity, Aspect, Polarity) triplet generated by the models. The following table presents model-level breakdowns across the three entity categories.

Model	Actor			Location			Event		
	Cor.	Inc.	Acc(%)	Cor.	Inc.	Acc(%)	Cor.	Inc.	Acc(%)
GPT	11	3	78,57	11	3	78,57	14	0	100,
Mistral-large	34	3	91,89	39	0	100,	44	3	93,62
Mistral-nemo	30	5	85,71	18	11	62,07	18	7	72,
Llama-3.1	18	1	94,74	18	5	78,26	36	3	92,31
Gemeni-pro	19	11	63,33	13	5	72,22	39	12	76,47

TABLE 3. English: LLM performance across Actor, Location, and Event tasks

GPT 4.0 GPT displayed mixed performance across entity categories. It effectively identified several key events—such as "damage to Baltic Sea gas pipeline", "drop in pressure", and "suspected sabotage"—but struggled with specific actors (e.g., "Finland's Prime Minister", "Chinese container ship") and key locations (e.g., "Baltic Sea", "Tamar", "Nord Stream pipeline"). Notable inconsistencies emerged across entity types, with frequent omissions and occasional misclassifications.

Mistral Large Mistral Large emerged as the top performer. It demonstrated strong, consistent performance across all entity categories, achieving high accuracy with minimal incorrect entity-aspect pairs. The model successfully captured specific, challenging

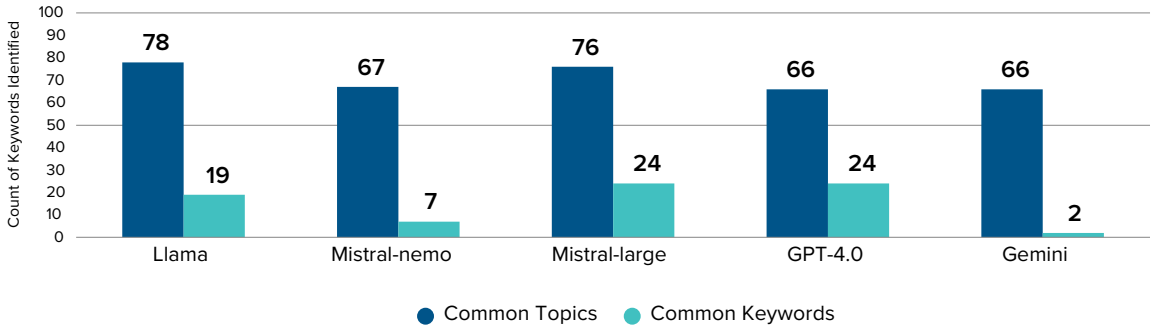

FIGURE 3. Visualising common topics and common keywords across LLMs

entities such as “Norsar”, “KRP”, and “Nord Stream pipeline”, along with less obvious actor mentions like “Detective Supt. Risto Lohi” and locations like “Paldiski”. It also inferred a broader and more coherent range of implicit aspects than other models.

Mistral-Nemo Mistral-nemo yielded moderate, and albeit inconsistent, results. It correctly identified core actors (e.g., “NATO”, “Finland’s Prime Minister”) and some events, but failed to recognise several specific locations and actors such as “Norsar”, “KRP”, and “Chinese container ship”. Its geographical coverage was uneven, with gaps in references such as the “Gulf of Finland”, reducing coherence across entity categories.

Llama-3.1 Llama-3.1 performed reliably, particularly in detecting event-related entities. It also identified locations such as “EEZ”, “Tamar”, and “Inkoo”, which other models overlooked. However, its performance in identifying actors and locations was more uneven. It showed occasional errors in classifying prominent locations like “Russia” or “Finland.”

In the evaluation dataset, Llama-3.1 successfully identified several specific locations missed by GPT and Mistral-Nemo (including ”EEZ”, ”Tamar”, and “Nord Stream’), but incorrectly flagged core locations as ’Finland’, ’Baltic Sea coast’, ’Russia’.

Actor identification was adequate but incomplete; it correctly identified “Gasgrid Finland” Strengths of Llama-3.1 included its strong performance in identifying diverse Event entities and some locations (such as ’Inkoo’ and ’Paldiski’), while its weaknesses centred on variable accuracy in actor and location recognition.

Gemini Pro Gemini Pro showed the weakest performance. It produced a high number of incorrect entity-aspect pairs across all categories. While it successfully identified some prominent events (e.g., “sudden drop in pressure”), it frequently failed to detect key actors (“KRP”, “Norsar”) and locations (“Baltic Sea coast”, “Mediterranean Sea”). Although it achieved high correct counts for certain event entities (e.g., ’Sudden drop in pressure’, ’Rise in gas prices’), these were undermined by similarly high error rates on central events (’suspected sabotage’, ’damage to Baltic-connector pipeline’). Its overall reliability was low, with frequent false positives, inconsistent inference, and unstable recognition.

7.4.2. Latvian

For Latvian, we also assessed the performance of five LLMs — GPT-4, Mistral Large, Mistral-nemo, Llama-3.1, and Gemini Pro — on the ABSA task in zero-shot settings. Latvian's complex morphology, inflectional grammar, and limited training data pose unique challenges for LLMs. As with English, we did not have the luxury of involving more evaluators for reviewing the outputs by the models. Instead, one author[3] manually reviewed the outputs to assess each (Entity, Aspect, Polarity) triplet generated by the models.

The following table presents model-level breakdowns across the three entity categories.

Model	Actor			Location			Event		
	Cor.	Inc.	Acc(%)	Cor.	Inc.	Acc(%)	Cor.	Inc.	Acc(%)
GPT-4	23	3	88,46	26	1	96,30	31	1	96,88
Mistral-large	41	10	80,39	59	8	88,06	123	73	62,76
Mistral-nemo	31	6	83,78	37	10	78,72	41	2	95,35
Llama-3.1	34	4	89,47	35	15	70,00	40	3	93,02
Gemeni-pro	63	5	92,65	80	14	85,11	54	8	87,10

Table 4. LLM performance across Actor, Location, and Event tasks

GPT 4.0 GPT-4 demonstrates the highest accuracy in ABSA for Latvian, excelling in all three categories — Location, Actor, and Event. GPT-4 achieved near-perfect accuracy in Location (96.30%) and Event (96.88%). Its ability to identify spatial references with only one error suggests robust handling of Latvian toponyms (e.g., "Riga" vs. "Daugavpils") and syntactic structures like locative cases (-ā, -ē). However, its Actor accuracy (88.46%, 3 errors) lagged slightly, likely due to challenges in disambiguating morphologically similar agentive nouns (e.g., "veikals" [store] vs. " pārdevējs" [seller] in Latvian). GPT-4 tends to identify fewer aspects than other LLMs, suggesting the model may be overly conservative, missing less prominent entities but maintaining high accuracy when it does commit to an aspect. Despite this, GPT-4 remains the most reliable model for the Latvian ABSA task, with an overall accuracy of approximately 94.12% (out of 85 identified aspects, 80 were correctly identified). Most of the errors produced by this model can be attributed to linguistic ambiguity. In all evaluated articles, a few abbreviations were misinterpreted, leading to incorrect classifications. Additionally, the model's tendency to focus on only a small portion of a paragraph – sometimes just a short phrase - often results in an incomplete contextual interpretation.

Mistral Large Mistral Large exhibited marked inconsistencies: while it performed moderately in Location (88.06%) and Actor (80.39%), its Event accuracy dropped sharply to 62.76% (73 errors) — the lowest among all models. Although the results show reasonable performance for Location and Actor entities, a substantial decline was observed for Event entities. Table 4 shows that the model recognised more events than other LLMs, yet accuracy decreased because the model applied the same explanatory structure — Excerpt, Explanation, Confidence Level, Polarity, and Polarity Score — to every entity, regardless of suitability. This resulted in a mixture of correct

3 Liene Zivere is a native speaker of Latvian as well as a Latvian linguist.

and incorrect outputs, often driven more by structural overgeneration than genuine interpretation. One article, demonstrated this issue most clearly, with similar patterns present in other articles, the scale of the problem was not as prominent in those cases.

Mistral-Nemo Mistral-Nemo performed very strongly in Event identification (95.35%), trailing only GPT-4. However, it lagged in Location (78.72%). Its balanced Actor performance (83.78%) places it in the middle of the cohort, but its high Event accuracy (41 correct, 2 errors) suggests strong capability at parsing dynamic scenarios. Comparatively, it outperformed Llama-3.1 in Location but fell short against Gemini Pro and GPT-4. This model demonstrated notable accuracy in extracting relevant excerpts that effectively conveyed meaning, making it easier for a human annotator to understand the intended message. This likely explains why it outperformed the other models in that section. However, some errors occurred when the model selected the same term for both the entity and the aspect. Additional inaccuracies also arose from incorrect aspect assignments, which reduced overall performance.

Llama 3.1 Llama 3.1 shows moderate performance in ABSA, with a noticeable gap between its accuracy and that of GPT-4. Llama-3.1 achieved the second-highest Actor accuracy (89.47%); however, it struggled with Location (70.00%), the lowest of all models in this category. One recurring error — consistent with the narrative detection task — is the model's tendency to confuse locations with actors when a term could plausibly function as both, indicating an uncertain grasp of article context. The model also selected contextually inappropriate aspects, such as environmental aspects despite no supporting evidence in the article.

Gemini Pro Gemini Pro demonstrated the most balanced performance, leading in Actor (92.65%) and maintaining strong Location (85.11%) and Event (87.10%) scores. It also stood out for using the shortest, yet precise excerpts, consistently selecting accurate and contextually grounded phrases. The primary cause of incorrect aspect classifications was the model's misidentification of polarity and polarity scores. In several cases, polarity could only be correctly interpreted by considering the full article context, which the model did not adequately account for.

7.5. Conclusion

The evaluation of LLMs on the English dataset revealed clear performance patterns across entity categories. Mistral Large emerged as the top performer, showing consistent and accurate identification across events, actors, and locations, successfully recognising both prominent and obscure entities such as "Norsar," "KRP," and "Paldiski." GPT-4.0 demonstrated strength in detecting key events like "damage to Baltic Sea gas pipeline" and "suspected sabotage" but struggled with identifying specific actors and locations, leading to inconsistent results. Mistral-Nemo showed moderate yet uneven performance; it correctly identified core actors like "NATO" and some events, but frequently missed specific entities such as "Chinese container ship" and "Gulf of Finland." Llama-3.1 performed reliably in detecting event-related entities and uniquely captured certain locations like "EEZ" and "Inkoo" that others overlooked. However, its performance in actor and location detection varied, and it sometimes misclassified important places like "Finland" and "Russia." Gemini Pro had the weakest performance overall, with high error rates, frequent false positives, and unreliable output despite correctly identifying some events. Overall, Mistral Large stood out for its broad and precise coverage, while Gemini Pro lagged behind significantly in reliability and accuracy.

In conclusion, for the Latvian dataset, while GPT-4 excels in accurately identifying Location and Event entities, Gemini Pro demonstrates superior accuracy in identifying Actors. However, no single model outperforms the others across all categories. Considering the

challenges of aspect identification in Latvian, event entities were the most consistently identified, achieving an average accuracy of 86.82%, while location entities presented the greatest difficulty, at 83.64%. This outcome suggests that the model had recognised locations in the article because only a few phrases explicitly encode toponyms, identical word forms were sometimes interpreted differently — for example, terms used to denote a country may also indicate the nationality of a prime minister. While the models performed adequately overall, errors arose when they misinterpreted sentence structure or contextual cues, leading to incorrect entity–aspect links. Future efforts should prioritise the curation of Latvian-specific training corpora and the fine-tuning of LLMs with these resources.

8. Key Findings and Implications

This study evaluated the performance of state-of-the-art LLMs on three strategic NLP tasks—narrative detection, topic modelling, and ABSA in both English and Latvian. The evaluation revealed both the potential and limitations of prompt-based LLM's in multilingual strategic communication contexts. In the narrative detection task, LLMs demonstrated promising capabilities in identifying key entities and relationships within complex texts. English-language outputs were often structurally coherent and followed the expected narrative layers (NER, RE, Plot Discovery, Story Evolution). However, zero-shot outputs still lacked the precision, contextual nuance, and inferential depth of human-generated analyses. In Latvian, model outputs were generally less complete, occasionally omitting salient entities or failing to construct coherent event chains. This disparity reflects the imbalance in training data availability across languages.

Topic modelling proved to be the most robust of the three tasks. All evaluated models produced coherent topic clusters in English, with consistent key terms and interpretable themes. The LLM-based prompt approach enabled unsupervised topic extraction without requiring pre-trained classifiers or fine-tuning. In Latvian, performance was slightly weaker, but still operationally useful, particularly when human post-editing refined noisier outputs.

ABSA was the most challenging task. LLMs struggled to distinguish reliably between sentiment targets and sentiment polarity in both English and Latvian. In zero-shot mode, they often relied on generic sentiment cues and failed to align specific sentiments with their respective aspects. Few-shot prompting improved performance marginally in English but had limited or negligible impact in Latvian.

Taken together, these results highlight the substantial capabilities of modern LLMs for strategic communication tasks, particularly in English, while also revealing persistent challenges in underrepresented EU languages such as Latvian. The most significant gains were achieved through carefully structured prompt design, suggesting that interface-level optimisations can partially compensate for the lack of language-specific fine-tuning. While English outputs generally showed higher fluency, coherence, and structural accuracy, they are not yet sufficient to replace human annotators — especially in zero-shot learning settings. Nonetheless, LLMs can significantly support human analysts by accelerating entity extraction, generating candidate relationships, and proposing narrative structures that guide the annotation process. In Latvian, LLMs offer useful but limited assistance and require even closer oversight, given occasional omissions or misinterpretations.

9. Repository and Resources

To support transparency and reproducibility, all materials used in this study are made publicly available in a dedicated GitHub repository[34].

The repository includes:

- Source code and evaluation scripts for all three tasks: Narrative Detection, Topic Modelling, and Aspect-Based Sentiment Analysis (ABSA).
- Input corpora in English and Latvian, as used in the evaluation.
- Prompt templates tailored for different LLMs to ensure consistent and structured task execution.
- A README file describing the project structure, usage instructions, and contributor information.

This resource is intended to facilitate further experimentation and comparative analysis in multilingual NLP tasks, particularly for low-resource languages.

Endnotes

1 Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali and Monojit Choudhury, 'The State and Fate of Linguistic Diversity and Inclusion in the NLP World', in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (2020), 6282–6293.

2 SoimulPatriei, *'stratcom-llm-evaluation'*, GitHub repository, [Accessed 4 December 2025.

3 Laria Reynolds and Kyle McDonell, 'Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm', *arXiv* preprint, arXiv:2102.07350 (2021).

4 OpenAI, 'GPT-4 Technical Report', *arXiv* preprint, arXiv:2303.08774 (2023).

5 Microsoft, *'Azure OpenAI Service Documentation'*, Microsoft Learn, [Accessed 4 December 2025].

6 Mistral AI, 'Mistral Models and Inference', technical documentation (2023).

7 OpenRouter, *'API Reference Overview'*, [Accessed 4 December 2025].

8 Mistral AI, 'Mistral Models and Inference', technical documentation (2023).

9 Meta AI, *'Llama 3: Open Foundation and Instruction Models'*, technical report (2024), [Accessed 4 December 2025].

10 Google DeepMind, *'Gemini: Google's Multimodal LLM'*, technical report (2023), [Accessed 4 December 2025]

11 Euronews, *'Finland Investigating Possible Sabotage of Baltic Gas Pipeline to Estonia'*, 8 October 2023, [Accessed 4 December 2025].

12 BBC News, *'Finland Investigates Suspected Sabotage of Baltic-Connector Gas Pipeline'*, 10 October 2023, [Accessed 4 December 2025].

13 Associated Press, *'Damage to Gas Pipeline, Telecom Cable Connecting Finland and Estonia Caused by "External Activity", Finland Says'*, AP News, 10 October 2023, [Accessed 4 December 2025].

14 Reuters, *'Kremlin, Asked about Damaged Baltic Pipeline, Says Threats to Russia "Unacceptable"'*, 23 October 2023, [Accessed 4 December 2025].

15 The Asahi Shimbun, *'Norwegian Navy Shadows Chinese Vessel Probed over Baltic Pipe Damage'*, 18 October 2023, [Accessed 4 December 2025].

16 The Straits Times (via AFP), 'China, Finland Held "Constructive" Talks on Damaged Gas Pipeline', 10 January 2024.

17 South China Morning Post, *'Chinese and Finnish Presidents Discuss Damage to Baltic Gas Pipeline Blamed on Chinese Ship'*, 10 January 2024, [Accessed 4 December 2025].

18 Eduard Barbu, Somnath Banerjee, Marija Isupova and Yukai Zeng, *Narrative Detection and Topic Modelling in the Baltics* (Riga: NATO Strategic Communications Centre of Excellence, 2024), 34 pp. [Accessed May 2024].

19 Inderjeet Mani, *Computational Modeling of Narrative* (San Rafael, CA: Morgan & Claypool, 2012).

20 Pratik Ranade, Sohan Dey, Ashwin Joshi and Tim Finin, 'Computational Understanding of Narratives: A Survey', *IEEE Access* 10 (2022): 101575–101594.

21 Vikas Yadav and Steven Bethard, 'A Survey on Recent Advances in Named Entity Recognition from Deep Learning Models', in *Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018)*, ed. Emily M. Bender, Leon Derczynski and Pierre Isabelle (2018), 2145–2158.

22 Makoto Miwa and Mohit Bansal, 'End-to-End Relation Extraction Using LSTMs on Sequences and Tree Structures', in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (ACL 2016), 1105–1116.

23 David M. Blei, Andrew Y. Ng and Michael I. Jordan, 'Latent Dirichlet Allocation', *Journal of Machine Learning Research* 3 (2003): 993–1022..

24 Dimo Angelov, 'Top2Vec: Distributed Representations of Topics', *arXiv* preprint, arXiv:2008.09470 (2020).

25 Maarten Grootendorst, 'BERTopic: Neural Topic Modeling with Class-Based TF-IDF and BERT Embeddings', *arXiv* preprint, arXiv:2203.05794 (2022).

26 M. Reuter, A. Thielmann, C. Weisser, S. Fischer and B. Safken, 'GPTopic: Dynamic and Interactive Topic Representations', preprint (2024).

27 X. Ding, J. Zhou, L. Dou, Q. Chen, Y. Wu, A. Chen and L. He, ***'Boosting Large Language Models with Continual Learning for Aspect-Based Sentiment Analysis'***, in *Findings of the Association for Computational Linguistics: EMNLP 2024* (2024), 4367–4377.

28 R. Fan, S. Li, T. He and Y. Liu, 'Aspect-Based Sentiment Analysis with Syntax-Opinion-Sentiment Reasoning Chain', in *Proceedings of the 31st International Conference on Computational Linguistics (COLING 2025)*, ed. O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio and S. Schockaert (2025), 3123–3137.

29 W. Zhang, Y. Deng, B. Liu, S. Pan and L. Bing, 'Sentiment Analysis in the Era of Large Language Models: A Reality Check', in *Findings of the Association for Computational Linguistics: NAACL 2024* (2024), 3881–3906.

30 B. Ozyurt and M. A. Akcayol, 'A New Topic Modeling-Based Approach for Aspect Extraction in Aspect-Based Sentiment Analysis: SS-LDA', *Expert Systems with Applications* 168 (2021): 114231.

31 J. Šmíd and P. Král, 'Cross-Lingual Aspect-Based Sentiment Analysis: A Survey on Tasks, Approaches, and Challenges', *Information Fusion* 120 (2025).

32 W. Zhao, Z. Yang, S. Yu, S. Zhu and L. Li, 'Contrastive Pre-Training and Instruction Tuning for Cross-Lingual Aspect-Based Sentiment Analysis', *Applied Intelligence* 55, no. 5 (2025).

33 E. Jang, J. Cui, D. Yim, Y. Jin, J.-W. Chung, S. Shin and Y. Lee, 'Ignore Me but Don't Replace Me: Utilizing Non-Linguistic Elements for Pretraining on the Cybersecurity Domain', in *Findings of the Association for Computational Linguistics: NAACL 2024* (2024), 29–42.

34 SoimulPatriei, ***'stratcom-llm-evaluation'***, GitHub repository, [Accessed 4 December 2025].

Prepared and published by the
NATO STRATEGIC COMMUNICATIONS
CENTRE OF EXCELLENCE

The NATO Strategic Communications Centre of Excellence (NATO StratCom COE) is a NATO accredited multi-national organisation that conducts research, publishes studies, and provides strategic communications training for government and military personnel. Our mission is to make a positive contribution to Alliance's understanding of strategic communications and to facilitate accurate, appropriate, and timely communication among its members as objectives and roles emerge and evolve in the rapidly changing information environment.

www.stratcomcoe.org | @stratcomcoe | info@stratcomcoe.org