Llama

While LLM demonstrates notable strengths in named entity recognition and plot discovery, it struggles with accurate entity classification, relationship extraction, and story evolution. The model’s tendency to simplify relationships and overlook nuanced shifts in narrative structure limits its ability to fully capture an article’s depth.

In contrast, human annotators consistently outperform the model in these areas due to their capacity for recognizing complex interconnections and implicit themes. Future improvements in AI model should focus on refining contextual understanding, improving entity classification, and enhancing the ability to track evolving narratives over time.

**Named Entity Recognition**

Model showed mixed performance in Named Entity Recognition. In the Delfi article, it correctly identified "somu jūras līcis" as a location but missed key actors and events. It also struggled with timeline accuracy, incorrectly adding an extraneous date.

In the LA article, the model misclassified NATO and the EU as locations instead of actors, missed key figures, and failed to recognize some timelines. Similar issues appeared in the LSM article, though it showed slight improvement in identifying timelines.

For the TVNET article, actor recognition improved, but the model continued to misclassify NATO as a location. It also overemphasized certain events while struggling with timeline, limiting its ability to capture the full timeline of events.

**Relationship Extraction**

Model for relationship extraction was generally correct but lacked depth. In the Delfi article, it linked entities but only in an obvious manner without deeper insights. The LA article showed an attempt at deeper understanding, but the model failed to construct meaningful connections beyond surface-level relationships.

For the LSM article, relationships were accurate but incomplete, with the model oversimplifying explanations and omitting key connections. The TVNET article showed improvement, identifying more relationships, though some were incorrectly tagged, highlighting inconsistency in its approach.

**Plot Discovery**

Across all articles, model performed relatively well in plot discovery. Although it occasionally assigned events differently than human annotators, it maintained a coherent representation of the story's structure. However, a recurring issue was its tendency to condense information into a highly summarized form, sometimes missing relevant details. For instance, in the Delfi article, while it successfully captured the story's essence, its presentation lacked the depth found in human annotations.

Similarly, in the LA article, the model provided a strong plot summary but failed to extract details beyond the climax. This issue was also observed in the LSM article, where it focused heavily on NATO and EU discussions while neglecting other key aspects of the resolution.

**Story Evolution**

Story evolution proved to be one of models weaker areas. In multiple articles, the model focused on only one major shift while failing to capture additional narrative developments. In the Delfi article, for example, it centered on a single aspect of the story while neglecting the broader evolution of events. A similar issue was observed in the LA article, where it omitted crucial details contributing to the story’s progression.

In the LSM article, model again overemphasized NATO and EU-related content in its analysis of story evolution. This focus detracted from a more balanced understanding of the article’s progression.

Delfi

#### **Named Entity Recognition (NER)**

Model correctly identified *“somu jūras līcis”* as a location, aligning with the human annotator’s assessment. However, while its overall performance was consistent with human annotation, it failed to recognize all actors, including the most crucial one for the article.

In terms of event recognition, the model performed poorly, identifying only a limited number of events and failing to provide a comprehensive explanation of the occurrences described in the article. Additionally, its timeline recognition was inconsistent, as it correctly identified only two dates but also incorrectly included a date, *October 8*.

#### **Relationship Extraction**

Model demonstrated an acceptable performance in relationship extraction by correctly associating certain entities. However, the identified relationships lacked depth, presenting only surface-level correlations rather than a more nuanced understanding of their interactions. For example, while it recognized the connection between *Balticconnector*, Finland, and Estonia, this association was an obvious one rather than a meaningful analytical insight.

#### **Plot Discovery**

The model performed relatively well in plot discovery, but with some differences in how it categorized key elements of the story. While its structuring of events differed from that of a human annotator, it successfully captured the overall narrative. Additionally, model presented its findings in a highly condensed manner, extracting only the most critical aspects of the story.

#### **Story Evolution**

Model focused on only a single aspect of the article, overlooking other key developments. While its interpretation was not entirely incorrect, this narrow focus prevented it from capturing the full scope of the story’s progression and the various narrative shifts that occurred. Consequently, its analysis was insufficient for a comprehensive understanding of the story’s evolution.

—

LA

**Named Entity Recognition (NER)**

Model incorrectly classified EU and NATO as locations rather than actors. Additionally, it failed to recognize some locations and occasionally provided excessively detailed names, such as “Inčukalma gāzes krātuve” instead of the more commonly used “Inčukalns.” While not incorrect, this demonstrates a deviation from conventional references.

Regarding actor recognition, the model identified some actors but failed to include their titles, reducing clarity. Additionally, it omitted certain key actors, such as Somijas robežsardze. In terms of event recognition, while it detected some occurrences, it overlooked related incidents that, although not explicitly mentioned, were contextually relevant to the article. Furthermore, the model only recognized explicit date references while failing to identify day-based temporal expressions.

**Relationship Extraction**

Model recognized some relationships that were not explicitly stated in the article. While it successfully extracted some connections, it struggled to accurately associate entities in a meaningful way. Additionally, it exhibited similar limitations as in previous cases, failing to construct deeper, more contextually relevant relationships.

**Plot Discovery**

The model performed well in plot discovery, effectively capturing the overall narrative. While it omitted certain post-climax details that were present in the article, its summarization of the story remained coherent and well-structured.

**Story Evolution**

Model struggled with story evolution, failing to extract crucial details necessary for understanding the progression of events. Its shift detection was inaccurate, as it focused primarily on the main storyline rather than identifying narrative developments. Additionally, in causal relationship extraction, the model introduced elements that were not present in the article, further undermining the accuracy of its analysis.

—

LSM

**Named Entity Recognition (NER)**

Model once again misclassified NATO and EU as locations while also identifying Latvia as a location, despite its limited relevance in the article. Interestingly, the model recognized EU and NATO as actors in the evaluation phase, yet still categorized them as locations, indicating inconsistency in entity classification.

Regarding actor recognition, the model identified certain actors, including general references such as “institūciju pārstāvji” (institution representatives) and “ūdenslīdēji” (divers). However, it failed to comprehensively recognize all relevant entities, demonstrating inconsistency in its approach. Additionally, while the model provided a concise summary of recognized entities, it omitted significant details.

A notable improvement was its recognition of temporal expressions, successfully identifying “ceturtdiena” (Thursday) and “nākamā nedēļa” (next week), which had been a weakness in previous analyses.

**Relationship Extraction**

Model approached relationship extraction differently in this article. While the relationships it identified were generally accurate, the model failed to capture the full range of connections present in the text. Additionally, its tendency to shorten explanations left room for ambiguity, making some relationships less clear.

**Plot Discovery**

The model provided a generally coherent plot summary but attempted to make it as brief as possible, resulting in the omission of key details. In particular, its focus on NATO and EU in the falling action and resolution stages overlooked other significant aspects of the article that could have been highlighted.

**Story Evolution**

Model disproportionately focused on the final section of the article, particularly its discussion of NATO and EU, rather than capturing the broader evolution of the story. This narrow focus resulted in an incomplete representation of the article's narrative progression.

TVNET

**Named Entity Recognition (NER)**

ChatGPT once again misclassified NATO as a location, but aside from this, its entity recognition was largely consistent with human annotation. The model successfully identified the same actors as previous analyses but also inferred the involvement of national governments, despite their absence in the article.

Regarding event recognition, the model identified several key events but overemphasized the cable damage, mentioning it three times. Additionally, it failed to recognize timeline such as “dažu dienu laikā” (within a few days), demonstrating a continued struggle with more complex time-related references.

**Relationship Extraction**

The model correctly identified many significant relationships, though the way it assigned them was occasionally inaccurate. Despite these misclassifications, its performance showed improvement over previous analyses, capturing a greater number of relevant connections.

**Plot Discovery**

Model performed well in summarizing the article’s plot, though it placed particular emphasis on a different aspect of the falling action compared to previous cases. Overall, the model effectively covered the full scope of the article, providing a coherent and comprehensible summary.

**Story Evolution**

The model demonstrated reasonable accuracy in Shift Detection, aligning closely with human annotation. However, it continued its pattern of focusing heavily on the final section of the article, leading to an imbalanced representation of the story’s progression. While the extracted information was not incorrect, it lacked a broader perspective on how the narrative evolved throughout the text.