NOVELSUM: Evaluating Long-Form Summary Generation for Historical Scandinavian Novels

Ali Al-Laith[1], **Alexander Conroy**[1], **Kirstine Nielsen Degn**[1], **Jens Bjerring-Hansen**[1] **and Daniel Hershcovich**[2]
[1] Department of Nordic Studies and Linguistics, University of Copenhagen, Denmark
[2] Department of Computer Science, University of Copenhagen, Denmark
{alal, dh}@di.ku.dk
{alc, knd, jbh}@hum.ku.dk

Abstract

We study long-form summarization of late-19th-century Danish and Norwegian novels and propose NOVELSUM, an evaluation resource and protocol tailored to literary narrative. We use a curated set of historical novels paired with professional reference summaries to establish baselines with long-document encoder–decoder models and prompt-based large-context LLMs. We evaluate with automatic metrics, expert human judgments, and LLM-as-judge scoring. Our human study identifies evaluation dimensions and literary facets that achieve substantial inter-annotator agreement and align with scholarly expectations. We further analyze reference-free evaluation, showing when it tracks expert trends and where it fails (notably for factual and setting-related criteria), thereby clarifying its utility when gold references or expert readers are unavailable. Our results benchmark long-context and prompted LLM approaches on historical literary prose and offer a practical path for human-grounded and reference-free assessment.

Keywords: Text Summarization, Large Language Models, Historical Text, Novels, Digital Humanities

1. Introduction

Automatic summarization of long-form literary narratives, especially novels, remains a core open problem in NLP. Unlike news or encyclopedic prose, novels feature multi-threaded plots, evolving characters, and multiple, sometimes conflicting, discourses that unfold across hundreds of pages, making content selection and faithfulness particularly challenging (Fabbri et al., 2020). Progress here would benefit digital humanities at scale and help scholars engage with large bodies of understudied literature.

We target late-19th-century Danish and Norwegian novels, a setting that adds diachronic complexity (orthographic shifts, archaic vocabulary, long periodic sentences) and thus stresses contemporary summarizers. Building on BOOKSUM's end-to-end framing for literary summarization (Kryściński et al., 2021), we draw from a curated subset of Scandinavian novels (Bjerring-Hansen et al., 2022) paired with human-written reference summaries and adapt long-context abstractive methods to this historical domain.

Methodologically, we establish baselines with long-document encoder–decoder models, and compare them to prompt-based summarization with contemporary large-context LLMs. We evaluate using both automatic metrics and expert human judgments, and we probe reference-free evaluation to mitigate reliance on copyrighted references and expert readers.

Our contributions are twofold: (i) a new evaluation resource and protocol for long-text summarization in historical Danish/Norwegian literature, and (ii) a systematic comparison of long-context baselines and LLM prompting, including a study of human criteria and reference-free automatic assessment in this literary domain.

Research questions. (1) Which evaluation dimensions for novel summarization yield high inter-annotator agreement and face validity for literary scholars? (2) To what extent can we automate summary evaluation without requiring human experts who have read the full novels? (3) How accurately and consistently can reference-free methods assess summary quality in the absence of gold reference summaries?

All pre-processing, training, and evaluation code is released under an open-source license in this repository: `https://github.com/mime-memo/NOVELSUM`

2. Related Work

2.1. Text Summarization

Text summarization methods are generally classified as extractive or abstractive. Extractive approaches select salient sentences from the source text, often producing summaries that may lack coherence and natural flow. In contrast, abstractive methods generate new sentences by paraphrasing and condensing content, resulting in more human-like summaries but requiring stronger language un-

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 9953–9963
11–16 May 2026.

derstanding and generation capabilities. One extractive approach employs topic modeling to identify candidate sentences associated with key topic words in long, unstructured novel texts, then selects the most important and diverse sentences to balance compression, quality, and readability, followed by a smoothing step to improve coherence (Wu et al., 2017).

Recent advances in deep learning and pretrained language models have greatly enhanced abstractive summarization, though both paradigms retain distinct strengths and limitations (Chen et al., 2019; Giarelis et al., 2023; Kirmani et al., 2024). The LOCOST model, an encoder–decoder architecture based on state-space models, efficiently handles very long input sequences and achieves performance comparable to sparse transformers while using far less memory. It can summarize texts exceeding 600,000 tokens, setting new benchmarks for full-book summarization (Bronnec et al., 2024).

Large language models (LLMs) such as GPT-4, Claude, and Llama now enable the summarization of long documents using extended context windows and hierarchical or recursive strategies. These models can process millions of tokens, making them suitable for book-length summarization (Chang et al., 2023). However, maintaining coherence and factual consistency across long contexts remains challenging, and chunking or hierarchical merging is often employed to mitigate context limitations (Wang et al., 2023). Several models were introduced to perform automatic abstractive summarization in Danish (Kolding et al., 2023). These models were fine-tuned using an abstractive subset of the DaNewsroom dataset (Varab and Schluter, 2020).

2.2. Scandinavian Historical and Literary Text Corpora

Processing historical texts presents unique challenges, including non-standardized orthography, archaic vocabulary, and complex sentence structures (Romary, 2012). These factors complicate tasks such as tokenization, part-of-speech tagging, and normalization. LLMs and other NLP tools often struggle with the variability and diachronic changes in historical languages, necessitating specialized preprocessing and normalization techniques (Schoffel et al., 2025)

Several resources exist for historical Nordic languages, including corpora of Danish, Swedish, and Norwegian (Bjerring-Hansen et al., 2022; Pettersson and Borin, 2022; Heinsen and Bøgeskov, 2025). These resources often provide transcriptions, translations, and facsimiles, supporting both linguistic and literary research. However, access to high-quality, digitized, and annotated corpora remains a challenge, and researchers have begun to turn their attention to the limited but high-quality data produced for traditional scholarly editions (Baunvig and Rasmussen, 2025; Rasmussen and Vad, 2025; Nimb, 2025; Conroy, 2025).

The development of benchmark datasets for historical and literary texts has gained significant momentum in recent years, reflecting the growing interest in evaluating large language models (LLMs) on nuanced linguistic and cultural phenomena. Recent efforts have produced datasets targeting diverse tasks, including sentiment analysis (Al-Laith et al., 2023), named entity recognition (Pettersson et al., 2024), semantic change (Kutuzov et al., 2022), word sense disambiguation (Al-Laith et al., 2024a), euphemism detection (Al-Laith et al., 2025), metaphor explanation (Pedersen et al., 2025), and noise and sound categorization (Al-Laith et al., 2024b), collectively enriching the evaluation landscape for LLMs in complex literary contexts.

2.3. Evaluation in Summarization

There are different metrics to perform the automatic evaluation of summarization. ROUGE remains the most widely used metric for summarization evaluation, though it is limited by its reliance on lexical overlap and struggles with abstractive outputs (Barbella and Tortora, 2022). BLEU, BERTScore, and newer metrics such as MoverScore and QuestEval have been proposed to better capture semantic similarity and factual consistency (Scialom et al., 2021; Fabbri et al., 2020). Embedding-based and reference-free metrics are increasingly adopted to address the shortcomings of traditional approaches.

Faithfulness and hallucination are critical concerns in neural summarization, especially for abstractive and multilingual settings. Metrics such as mFACT and entailment-based measures have been developed to assess factual consistency, with human evaluation often revealing substantial hallucination in model-generated summaries (Qiu et al., 2023; Maynez et al., 2020).

Human evaluation remains essential for assessing fluency, informativeness, coherence, and cultural adequacy. Protocols increasingly emphasize fine-grained annotation, error taxonomies, and multi-dimensional scoring to capture the nuances of summary quality, especially for long or complex texts (Fabbri et al., 2020; Maynez et al., 2020; Chang et al., 2023).

Although several Scandinavian NLP resources exist, they mostly focus on short or contemporary texts and linguistic tasks. Long-form summarization of older novels remains largely unexplored. This work addresses that gap by evaluating large language models on summarizing 19th-century literary

texts using both automatic and human evaluation criteria.

3. Dataset

We use the MeMo corpus (Bjerring-Hansen et al., 2022), a collection of 859 Danish and Norwegian novels (64M+ tokens) from the last 30 years of the 19th century[1]. It should be noted that until 1907, written Norwegian was practically identical to written Danish (Vikør, 2022). We use 34 novels from the MeMo corpus to perform the summarization experiments. The reference summaries of the selected novels were obtained from *Litteraturen i Danmark og de øvrige nordiske Lande: Hvem skrev hvad før 1914*, edited by Henning Fonsmark, a rigorous and highly valuable scholarly handbook of Danish literature (Fonsmark, 1961). The book remains copyrighted and contains 112 summaries of novels; 34 of these novels are included in the corpus. We compare these reference summaries with the automatically generated summaries in our experiments. Table 1 shows statistics from both selected novels and their summaries in our dataset.

Statistic	**Novels (avg.)**	**Reference Summaries (avg.)**
Word Count	64,712	246
Sentence Count	3,588	11
Token Count	77,985	288
Unique Words	9,000	159
AVG. Sentence Length	19.21	22.44

Table 1: Average linguistic statistics for the novels and their reference summaries in the dataset.

4. Methodology

4.1. Summarization-specific Models

For the baseline experiments, we employed three transformer-based models specifically designed for long-document summarization. The models were trained or fine-tuned on diverse datasets to efficiently handle long contexts and maintain coherence across extended narratives. These architectures incorporate specialized attention mechanisms—such as sparse or global attention—to reduce computational costs while preserving contextual understanding. They serve as strong abstractive baselines, providing a point of comparison for evaluating the performance of large language models under different prompting strategies used in our experiments.

[1]Released under the Creative Commons Attribution 4.0 license: https://huggingface.co/datasets/MiMe-MeMo/Corpus-v1.1.

4.2. Prompting-based Summarization

We use large language models (LLMs) that have not been specifically trained for summarization, by prompting them with instructions for the task. We use three prompting strategies: zero-shot, either with the full novel text in the prompt or hierarchically, and metadata-based, without the novel text in the prompt.

The first strategy applied *zero-shot summarization* when the entire text fit within the model's context window. If the novel's token count exceeded the model's maximum input length, which occurred in only a few cases, the text is divided into chunks matching the model's context window, and each chunk was summarized separately. In this setup, the model is directly prompted to produce a Danish summary of the novel, emphasizing essential narrative elements such as the fictional time, main places, central characters, and key plot events. The prompt explicitly instructed the model to focus on factual details rather than interpretation, encouraging concise and objective summaries. This approach allowed for a direct evaluation of each model's intrinsic summarization ability without additional context engineering.

The second strategy is *zero-shot hierarchical summarization*. Each novel was divided into paragraph chunks of approximately 30,000 tokens, which were summarized independently using the same zero-shot prompt. We apply structure-aware paragraph chunking with a maximum token constraint. The resulting partial summaries were then merged through an additional summarization step to form a coherent final summary. This hierarchical setup ensured that even very long texts could be processed effectively while maintaining consistency across segments and minimizing information loss due to truncation.

The third prompting setup leveraged *metadata-based summarization*. Instead of providing the full text, the model was given only the novel's title and author name and was asked to produce a factual Danish summary covering time, place, characters, and plot events. This strategy evaluated the model's parametric knowledge and its ability to generate summaries based on learned literary information rather than direct textual input. Comparing these metadata-based summaries with text-based ones provided insight into each model's background knowledge of historical Scandinavian literature and its ability to generalize beyond the provided text.

5. Experiments

5.1. Large Language Models

We employ different models for baseline and experimental setups. For baseline comparisons, we include three transformer-based long-document summarization models: **Longformer Encoder-Decoder (LED)**[2] (Beltagy et al., 2020), **LongT5-TGlobal-Base**[3] (Guo et al., 2021), and **DanSumT5-large**[4] (Kolding et al., 2023; Varab and Schluter, 2020). LED and LongT5 extend transformer architectures with sparse and transient-global attention to efficiently process long contexts, while DanSumT5-large is fine-tuned on the Danish DaNewsroom dataset for monolingual abstractive summarization.

For zero-shot and metadata-based experiments, we select one representative model from each of the four best-performing families on the Danish EuroEval leaderboard[5] . These include **DeepSeek-V3** (DeepSeek-AI, 2024), a Mixture-of-Experts model supporting 128K tokens; **Llama-3.2-3B-Instruct-Turbo**[6] , optimized for efficient instruction following; **Gemma-3n-E4B-it** (Team, 2025a), a lightweight multimodal model for constrained environments; and **GPT-4o-mini**, a compact variant of GPT-4 with strong reasoning and summarization abilities.

For model-based evaluation, we use **Qwen3-235B-A22B-Instruct-2507-FP8**[7] (Team, 2025b), a 235B-parameter instruction-tuned model known for high evaluation accuracy and efficiency using FP8 inference precision.

5.2. Baseline Summarization

The baseline summarization experiment was conducted using the three summarization-specific models introduced in Section 5.1. These models were selected to establish strong abstractive baselines due to their ability to process extended contexts through specialized attention mechanisms. Each model was evaluated on 34 novels to provide a comparative benchmark for assessing the performance of prompt-based large language models used in subsequent experiments.

[2] https://huggingface.co/allenai/led-base-16384

[3] https://huggingface.co/google/long-t5-tglobal-base

[4] https://huggingface.co/Danish-summarisation/DanSumT5-large

[5] https://euroeval.com/leaderboards/Monolingual/danish/

[6] https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct

[7] https://huggingface.co/Qwen/Qwen3-235B-A22B-Instruct-2507-FP8

5.3. Prompting Summarization

Different prompt variations are used to summarize historical novels using several LLMs. Below are the specific prompt structures used in the study:

Zero-shot Summarization Prompt In this setup, each model was prompted in a zero-shot manner to produce a Danish summary of the input novel without any prior examples or contextual guidance. The instruction explicitly asked the model to focus on factual narrative elements—such as fictional time, setting, main characters, and key plot events—while avoiding interpretation or stylistic elaboration. The system role framed the model as an attentive reader, and the user message contained the summarization request as follows:

> *"You are an attentive reader who is good at summarizing a long text. Please provide a Danish summary of the following novel. The summary should mention the fictional time, key places, main characters, and the main events of the plot. Focus on facts rather than interpretation:{text}"*

Zero-shot Hierarchical Summarization Prompt To handle longer novels exceeding the model's input length limit, we applied a zero-shot hierarchical summarization strategy. Each novel was first divided into chunks of approximately 30K tokens. The model was asked to summarize each chunk individually, and then a second-stage summarization merged these partial summaries into a coherent overall summary. In this setup, the model received multiple intermediate summaries and was instructed to synthesize them into a single, well-structured Danish summary without additional examples or guidance. The prompting template is similar to the Zero-shot summarization prompt; then summarize the summaries of each chunk as follows:

> *"You are an attentive reader who is good at summarizing a long text. Here are several summaries of different parts of a 19th-century Danish novel:{text}. Please merge these into a single, well-structured summary in Danish."*

Metadata-based Summarization Prompt In this setup, the model was not provided with the novel text but was instead prompted using only metadata — specifically, the novel's title and author name. The objective was to test whether large language models possess enough implicit literary and cultural knowledge to generate coherent and factually grounded summaries from prior training exposure.

The model was instructed to write a Danish summary that includes the fictional time, main characters, key places, and major plot events, focusing on factual accuracy rather than interpretation. The prompting template is as follows:

> *"You are an attentive reader who is good at summarizing a long text." Please provide a Danish summary of the following novel: '{novel_name}' written by: {author_name}. The summary should mention the fictional time, key places, main characters, and the main events of the plot. Focus on facts rather than interpretation."*

5.4. Evaluation

Automatic Evaluation We employ three metrics for automatic evaluation: 1) ROUGE (Lin, 2004) measures the lexical overlap of words or n-grams between generated text and a reference, emphasizing recall-based content similarity. 2) BERTScore (Zhang et al., 2019) evaluates semantic similarity by aligning contextual embeddings from a pretrained BERT model, allowing it to capture meaning beyond exact word matches. 3) Semantic Answer Similarity (SAS) (Risch et al., 2022) uses a fine-tuned cross-encoder to assess semantic equivalence between generated and reference summaries, making it particularly suitable for open-ended generation tasks such as summarization, where multiple semantically valid outputs may exist. For SAS, we use the MiniLM-L6-Danish-Encoder sentence transformer model[8].

Human Evaluation For our human evaluation experiments, literary scholars with domain expertise selected four novels: *Fru Marie Grubbe* (1876) and *Niels Lyhne* (1880) by J. P. Jacobsen, *Haabløse Slægter* (1880) by Herman Bang, and *Sult* (1890) by Knut Hamsun. It should be noted that these novels are highly canonical and frequently discussed in contemporary and later sources (such as encyclopedias, literary histories, and newspaper reviews), meaning that the models have likely been exposed both to the texts themselves and to second-hand sources discussing them. However, the reference summaries against which we evaluate the novels have never been digitized and are therefore unlikely to appear in model training data.

For the human evaluation, we provided detailed annotation guidelines covering both general linguistic quality and literary content. Three evaluators with a background in literature who had read the four novels assessed each generated summary along nine dimensions using a 5-point Likert scale (1 = poor, 5 = excellent) without access to the reference summaries, and the scores were averaged. First, standard summarization quality metrics were used: Fluency, Coherence, Relevance, and Factuality (Kryściński et al., 2021; Zhao et al., 2022).

Second, to capture the literary adequacy of summaries, annotators also evaluated how well the summaries preserved key narrative elements: Time, Place, Characters, Plot, and Themes. Although the scheme was developed to provide a general framework for summarizing novels, it can readily be tailored to address specialized analytical needs or particular literary features. These literary criteria reflect the ability of the summarization system to retain essential components of narrative structure, which are particularly relevant for historical Danish and Norwegian novels.

For the human evaluation experiments, we select four models with varying degrees of performance as reflected by the automatic evaluation results, to quantify the reliability of the automatic metrics over the full range of possible scores. The models are DeepSeek-V3, Llama-3.2-3B, gemma-3n-E4B-it and gpt-4o-mini, prompted in the same way as in the zero-shot summarization experiment (with the full novel text as input).

LLM as Judge To complement human evaluation and standard automatic metrics, we experimented with using LLMs as evaluators for summarization. While human judgments are reliable, they are time- and resource-intensive, and traditional metrics such as ROUGE or BERTScore may fail to capture narrative coherence, factual accuracy, or preservation of literary elements. We used Qwen3-235B-A22B-Instruct-2507-FP8 as the evaluation model, prompting it to act as an expert in literary analysis and score generated summaries on Fluency, Coherence, Relevance, and Factual Accuracy using a 5-point Likert scale. The evaluation prompts were carefully designed to mirror the human annotation guidelines, and deterministic decoding parameters (temperature = 0, top-p = 1) were used to ensure reproducibility. This approach allowed us to efficiently scale evaluation across large numbers of generated summaries while maintaining alignment with human judgments.

6. Results and Analysis

6.1. Automatic Evaluation Experiments

Table 2 presents the automatic evaluation results averaged across the 34 novels for the baseline and LLM-based summarization models across the three prompting setups. Among the baselines, LongT5-TGlobal-Base achieved the best overall per-

[8] https://huggingface.co/KennethTM/MiniLM-L6-danish-encoder

formance, surpassing LED and DanSumT5-large in all metrics. This confirms its stronger ability to model long-range dependencies in extended narrative texts.

Across the LLMs, GPT-4o-mini consistently produced the most accurate and semantically coherent summaries, achieving the highest BERTScore (0.6706) and strong ROUGE values in the zero-shot setting. DeepSeek-V3 followed closely, achieving the best SAS score (0.6067) and maintaining stable performance across all configurations. In the zero-shot hierarchical setup, performance slightly declined due to information compression during chunk merging; however, Llama-3.2-3B performed competitively on shorter segments, indicating robustness in handling divided input.

For the metadata-based summarization, where models were prompted only with the novel's title and author, the scores were generally lower but consistent. GPT-4o-mini again achieved the highest ROUGE results, while Gemma-3n-E4B-it obtained the best SAS score (0.5882), suggesting an ability to draw on implicit literary and narrative knowledge. Overall, GPT-4o-mini and DeepSeek-V3 emerged as the most reliable models, while smaller models offered more concise but less detailed summaries, highlighting the balance between efficiency and narrative richness.

6.2. Human Evaluation Experiments

The inter-annotator agreement (IAA) analysis using Krippendorff's α over the four selected novels and four models shows an overall reliability of 0.70, indicating a substantial level of agreement among the three human evaluators. Agreement scores varied across individual dimensions, with the highest consistency observed for Place (0.77), Time (0.74), and Relevance (0.73), reflecting strong alignment in judgments related to narrative setting and content relevance. Lower but still acceptable agreement values were found for Fluency (0.53) and Plot (0.59), suggesting greater subjectivity in assessing stylistic and narrative aspects. Overall, these results demonstrate reliable human evaluation performance across both linguistic and literary dimensions.

The human evaluation results indicate clear differences in summarization quality across models. Overall, DeepSeek-V3 and GPT-4o-mini produced the most fluent, coherent, and contextually accurate summaries, demonstrating stronger consistency in both standard and narrative dimensions. In contrast, gemma-3n-E4B-it and Llama-3.2-3B-Instruct-Turbo performed less effectively, particularly in capturing detailed narrative elements such as plot and setting. Table 3 shows details about human evaluation results using both standard and narrative metrics.

6.3. LLM-as-Judge Experiments

The model-as-judge evaluation on the 34 novels reveals clear differences in summarization quality across the tested models. Larger instruction-tuned models such as GPT-4o-mini and DeepSeek-V3 generally produced more fluent, coherent, and contextually faithful summaries, while smaller models tended to struggle with relevance and narrative completeness. Overall, the results show that model size and architectural sophistication contribute significantly to capturing key literary elements such as time, place, characters, and plot consistency in Danish novel summaries. Table 3 shows the details results of using Qwen3-235B-A22B-Instruct-2507-FP8 model as judge to evaluate the generated summaries compare to the reference summaries. We also conduct an experiment for the evaluation of generated summaries without providing the reference summaries as shown in Table 3. The results show that reference-free LLM-as-Judge evaluation is both accurate and consistent for assessing summary quality in the absence of gold references. It reliably distinguishes between strong and weak models and aligns well with human judgment trends, especially for fluency, coherence, and narrative elements. This makes it a viable and scalable alternative when reference summaries are unavailable or impractical to obtain.

6.4. Metric Reliability

We use Kendall's τ to calculate the correlation between human evaluation and both reference-free and reference-based automatic evaluations on standard and narrative metrics, over the four novels that have been selected for human evaluation. The results in Figure 1 show that both reference-based and reference-free evaluations are significantly correlated ($p < 0.05$) with human judgments across most metrics, indicating a strong overall alignment. However, in the reference-free setup, Factuality, Place, and Time did not show significant correlations, suggesting that these aspects remain challenging to assess accurately without access to gold reference summaries.

We also calculate the correlation between human evaluation scores and automatic evaluation metrics from the zero-shot summarization experiment, including SAS, ROUGE, ROUGE-L, and BERTScore. The results show that SAS exhibits the strongest and most consistent correlation with human judgments across both standard and narrative metrics, followed by BERTScore, which also aligns well with human assessments of fluency and coherence. In contrast, ROUGE-based metrics demonstrate weaker and mostly non-significant correlations, indicating that lexical overlap alone is insufficient to capture the qualitative and narrative

Model	BERTScore	ROUGE			SAS
		ROUGE-1	ROUGE-2	ROUGE-L	
Baseline Summarization					
LED	0.4524	0.0249	0.0004	0.0231	0.1180
LongT5	0.5981	0.1362	0.0047	0.1123	0.2193
DanSumT5-large	0.5909	0.1003	0.0018	0.0910	0.1408
Zero-shot Summarization					
DeepSeek-V3	0.6570	0.1742	0.0235	0.1608	**0.6067**
Llama-3.2-3B	0.6202	0.1581	0.0183	0.1471	0.4166
gemma-3n-E4B-it	0.6502	0.1541	0.0166	0.1423	0.4606
gpt-4o-mini	**0.6706**	0.1876	0.0226	0.1705	0.5865
Zero-shot Hierarchical Summarization					
DeepSeek-V3	0.6471	0.1509	0.0183	0.1397	0.5886
Llama-3.2-3B	0.6535	**0.1895**	0.0230	**0.1743**	0.5235
gemma-3n-E4B-it	0.6549	0.1599	0.0174	0.1466	0.4875
gpt-4o-mini	0.6701	0.1826	0.0207	0.1672	0.5908
Metadata-based Summarization					
DeepSeek-V3	0.6510	0.1650	0.0142	0.1533	0.5447
Llama-3.2-3B	0.6073	0.0806	0.0103	0.0763	0.4287
gemma-3n-E4B-it	0.6571	0.1627	0.0221	0.1495	0.5882
gpt-4o-mini	0.6612	0.1810	**0.0256**	0.1664	0.5850

Table 2: Automatic evaluation results for baseline and LLM models across three prompting setups, averaged over the 34 novels.

Model	Standard Metrics				Narrative Metrics					AVG.
	Fluency	Coherence	Relevance	Factuality	Time	Place	Characters	Plot	Themes	
Human Evaluation										
DeepSeek-V3	3.50	2.75	3.67	3.08	4.58	3.58	3.92	3.17	3.42	**3.52**
Llama-3.2-3B	1.83	2.42	1.42	2.42	1.92	2.00	2.17	1.92	1.58	1.96
gemma-3n-E4B-it	3.50	3.08	2.00	2.25	1.42	1.08	2.17	1.75	2.08	2.15
gpt-4o-mini	4.08	4.42	3.08	3.25	4.08	3.42	3.17	2.75	2.83	3.45
Reference-based LLM as a Judge										
DeepSeek-V3	5.00	5.0	4.50	3.75	4.5	4.50	4.00	4.00	4.25	**4.39**
Llama-3.2-3B	3.25	3.0	2.00	1.50	2.5	3.00	2.25	1.50	2.00	2.33
gemma-3n-E4B-it	5.00	5.0	2.50	1.50	2.5	2.25	2.25	1.75	2.25	2.78
gpt-4o-mini	5.00	5.0	4.25	3.25	4.5	4.50	3.75	3.75	4.50	4.28
Reference-free LLM as a Judge										
DeepSeek-V3	5.00	5.00	5.00	4.94	4.97	5.00	5.00	5.00	4.94	**4.98**
Llama-3.2-3B	3.79	3.24	2.97	2.82	3.38	4.09	3.47	2.82	2.94	3.28
gemma-3n-E4B-it	4.76	4.76	4.50	4.41	4.00	4.03	4.18	4.24	4.38	4.36
gpt-4o-mini	5.00	5.00	4.94	4.38	4.76	4.97	4.56	4.50	4.76	4.76

Table 3: Averaged evaluation results for human and LLM as a judge on standard and narrative metrics.

aspects of literary summarization. Overall, SAS and BERTScore emerge as more reliable indicators of human-perceived summary quality in this domain, as shown in Figure 2, almost (but not quite) on par with the reference-based LLM-as-a-Judge The latter remains the most reliable automatic metric to evaluate the summaries.

7. Conclusion and Future Work

We introduced NOVELSUM, an evaluation resource and protocol for long-form summarization of late-19th-century Danish and Norwegian novels, and benchmarked both long-context encoder–decoder models and prompted LLMs. Our study yields three main findings aligned with our research questions. (RQ1) We identify human evaluation dimensions—standard quality criteria plus literary facets (Time, Place, Characters, Plot, Themes)—that achieve substantial inter-annotator agreement and face validity for literary scholars, supporting reliable expert assessment in this domain. (RQ2) We show that automated evaluation can approximate expert trends without requiring readers who have studied the full novels: reference-based metrics (e.g., ROUGE, BERTScore, SAS) track human judgments on average, and LLM-as-judge correlates well with experts at the aggregate level. (RQ3) Reference-free evaluation is promising for fluency, coherence, and some narrative facets, but remains less dependable for factuality and set-

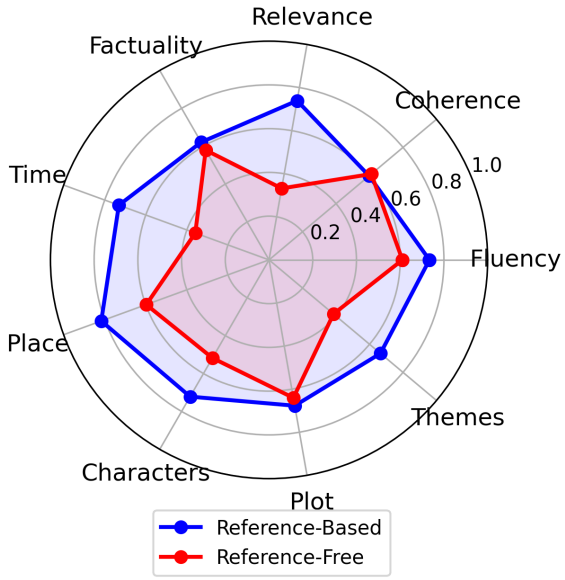

Figure 1: Human Evaluation Correlation with Reference-based and Reference-free LLM-as-a-Judge (Qwen3-235B-A22B-Instruct-2507-FP8) promoted for the corresponding aspects. Results are over 16 summaries (four novels and four models).

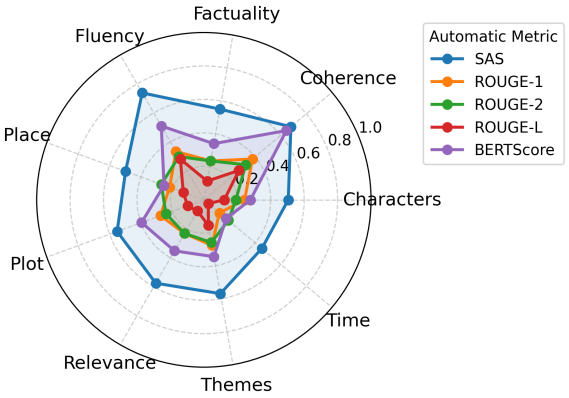

Figure 2: Human Evaluation Correlation with the Automatic Evaluation Metrics.

ting (Time/Place), clarifying when such methods are useful in the absence of gold references.

Future work will develop aspect-oriented summary generation and evaluation, explicitly prompting models for targeted facets (e.g., time, setting, character arcs, theme) and designing facet-specific rubrics and probes. This moves beyond static proxies toward situated, interpretable judgments about what a summary ought to include for a given cultural-literary context—an approach consistent with sociocultural accounts of positionality and indexicality, which model meaning as emergent from interaction rather than as trivia. Concretely, aspect prompts operationalize localization (what matters here, for this audience and text) and help disentangle when adaptation is appropriate from how it should manifest, aligning our agenda with calls to replace coarse cultural proxies with dynamic, theoretically grounded methods for cultural NLP (Zhou et al., 2025). Future work will also include expanding our empirical scope beyond the canon and incorporating some of the completely forgotten works from the corpus. This will allow us to dive deeper into the problem of memorization, as (in contrast to the canon) there is no information about these texts outside of our corpus.

Ethics Statement

Data sources and copyright. Our experiments use (i) the MeMo corpus of Danish and Norwegian novels (late 19th century), which is available under CC BY 4.0 and was accessed and processed in accordance with that licence; and (ii) short reference summaries printed in Litteraturen i Danmark og de øvrige nordiske Lande: Hvem skrev hvad før 1914 (ed. Henning B. Fonsmark). The MeMo novels themselves are public-domain texts; however, the book's summaries remain copyrighted (life+70) and the volume does not grant a reuse licence. We therefore treat those summaries as in-copyright material and do not redistribute them. All processing of copyrighted material was performed from lawfully obtained copies. Any internal copies created for analysis were used solely for research and are not redistributed.

Human evaluation. Three expert annotators (scholars of Scandinavian literature) voluntarily assessed model outputs using pre-specified guidelines. No personal, sensitive, or identifying information about the annotators or about individuals appears in the dataset; only aggregate scores are reported. The task involved literary quality judgements on model-generated text and posed minimal risk. According to our institutional guidance for non-personal-data studies, formal ethics board approval was not required.

Safety, bias, and responsible use. The resource concerns historical literary works; it does not include contemporary personal data or content targeting individuals or vulnerable groups. We report both automatic and human evaluations and caution against over-interpreting automatic metrics for long-form literary summarization. The released materials are intended solely for research and education.

8. Bibliographical References

Ali Al-Laith, Alexander Conroy, Jens Bjerring-Hansen, and Daniel Hershcovich. 2024a. Development and evaluation of pre-trained language models for historical Danish and Norwegian literary texts. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4811–4819, Torino, Italia. ELRA and ICCL.

Ali Al-Laith, Alexander Conroy, Jens Bjerring-Hansen, Bolette Pedersen, Carsten Levisen, and Daniel Hershcovich. 2025. Dying or departing? euphemism detection for death discourse in historical texts. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1353–1364, Abu Dhabi, UAE. Association for Computational Linguistics.

Ali Al-Laith, Kirstine Nielsen Degn, Alexander Conroy, Bolette Sandford Pedersen, Jens Bjerring-Hansen, and Daniel Hershcovich. 2023. Sentiment classification of historical Danish and Norwegian literary texts. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 324–334, Tórshavn, Faroe Islands. University of Tartu Library.

Ali Al-Laith, Daniel Hershcovich, Jens Bjerring-Hansen, Jakob Ingemann Parby, Alexander Conroy, and Timothy R Tangherlini. 2024b. Noise, novels, numbers. a framework for detecting and categorizing noise in Danish and Norwegian literature. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3344–3354, Miami, Florida, USA. Association for Computational Linguistics.

Marcello Barbella and G. Tortora. 2022. Rouge metric evaluation for text summarization techniques. *SSRN Electronic Journal*.

Katrine Frøkjær Baunvig and Krista Stinne Greve Rasmussen. 2025. Digitale udgaver som hitlgaranter. om editionsfaglig domæneekspertberigelse og datakvalitetens betydning for integrationen af computationelle tilgange i humanvidenskaben. In Annika Rockenberger, Aasta Marie Bjørvand Bjørkøy, Nina Marie Evensen, and Ellen Nessheim Wiger, editors, *Massedigitalisering og edisjonsfilologi. Bidrag til en konferanse arrangert av Nordisk nettverk for edisjonsfilologer*, volume 15 of *Nordisk nettverk for edisjonsfilologer. Skrifter*, pages 50–63. Universitetsbiblioteket i Oslo, Oslo.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Jens Bjerring-Hansen, Ross Deans Kristensen-McLachlan, Philip Diderichsen, and Dorte Haltrup Hansen. 2022. Mending fractured texts. a heuristic procedure for correcting ocr data. In *Proceedings of the 6th Digital Humanities in the Nordic and Baltic Countries Conference*, volume 3232, pages 177–186, Uppsala, Sweden. DHNB Proceedings.

Florian Le Bronnec, Song Duong, Mathieu Ravaut, Alexandre Allauzen, Nancy F Chen, Vincent Guigue, Alberto Lumbreras, Laure Soulier, and Patrick Gallinari. 2024. Locost: State-space models for long document abstractive summarization. *arXiv preprint arXiv:2401.17919*.

Yapei Chang, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2023. Booookscore: A systematic exploration of book-length summarization in the era of llms. *ArXiv*, abs/2310.00785.

Yangbin Chen, Yun Ma, Xudong Mao, and Qing Li. 2019. Multi-task learning for abstractive and extractive summarization. *Data Science and Engineering*, 4:14–23.

Carol J. Clover and John Lindow. 2019. Old norse-icelandic literature.

Alexander Conroy. 2025. Hvor langt skal vi fra teksten? om ordbøger, ai og verbalkommentarer. In Jeppe Barnwell, Simon Skovgaard Boeck, and Karen Skovgaard-Petersen, editors, *Ingen Kommentarer*, pages 111–128. Det Danske Sprog- og Litteraturselskab, Copenhagen.

DeepSeek-AI. 2024. Deepseek-v3 technical report.

A. R. Fabbri, Wojciech Kryscinski, Bryan McCann, R. Socher, and Dragomir R. Radev. 2020. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Henning B. Fonsmark. 1961. Litteraturen i danmark, og de øvrige nordiske lande: hvem skrev hvad før 1914.

Nikolaos Giarelis, Charalampos Mastrokostas, and N. Karacapilidis. 2023. Abstractive vs. extractive summarization: An experimental review. *Applied Sciences*.

Mandy Guo, Joshua Ainslie, David Uthus, Santiago Ontanon, Jianmo Ni, Yun-Hsuan Sung, and Yinfei Yang. 2021. Longt5: Efficient text-to-text transformer for long sequences. *arXiv preprint arXiv:2112.07916*.

Odd Einar Haugen, Massimiliano Bampi, M. Buzzoni, A. Meregalli, and L. Panieri. 2018. Le lingue nordiche nel medioevo.

Johan Heinsen and Camilla Bøgeskov. 2025. A world in print: Introducing a danish-norwegian corpus of historical newspapers. *arXiv preprint arXiv:2509.02356*.

Mahira Kirmani, Gagandeep Kaur, and Mudasir Mohd. 2024. Analysis of abstractive and extractive summarization methods. *Int. J. Emerg. Technol. Learn.*, 19:86–96.

Sara Kolding, Katrine Nymann, Ida Bang Hansen, Kenneth C Enevoldsen, and Ross Deans Kristensen-McLachlan. 2023. Dansumt5: Automatic abstractive summarization for danish. In *The 24rd Nordic Conference on Computational Linguistics*.

Wojciech Kryściński, Nazneen Rajani, Divyansh Agarwal, Caiming Xiong, and Dragomir Radev. 2021. Booksum: A collection of datasets for long-form narrative summarization. *arXiv preprint arXiv:2105.08209*.

Andrey Kutuzov, Samia Touileb, Petter Mæhlum, Tita Ranveig Enstad, and Alexandra Wittemann. 2022. Nordiachange: Diachronic semantic change dataset for norwegian. *arXiv preprint arXiv:2201.05123*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81.

ljubicamiocevic and CarolinOdebrecht. 2020. Swedish novel corpus (ELTeC-swe): Release with 58 novels.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan T. McDonald. 2020. On faithfulness and factuality in abstractive summarization. *ArXiv*, abs/2005.00661.

Sanni Nimb. 2025. Sprogteknologiske metoder til kommentering af digitale udgivelser – muligheder og begrænsninger. In Jeppe Barnwell, Simon Skovgaard Boeck, and Karen Skovgaard-Petersen, editors, *Ingen Kommentarer*, pages 83–109. Det Danske Sprog- og Litteraturselskab, Copenhagen.

Bolette S. Pedersen, Nathalie Sørensen, Sanni Nimb, Dorte Haltrup Hansen, Sussi Olsen, and Ali Al-Laith. 2025. Evaluating LLM-generated explanations of metaphors – a culture-sensitive study of Danish. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 470–479, Tallinn, Estonia. University of Tartu Library.

Lyonel Perabo. 2021. Analog resources and digital limitations. *Septentrio Conference Series*.

Eva Pettersson and Lars Borin. 2022. Swedish diachronic corpus.

Eva Pettersson, Lars Borin, and Erik Lenas. 2024. Swener-1800: A corpus for named entity recognition in 19th century swedish. In *Digital Humanities in the Nordic and Baltic Countries*, volume 6.

Yifu Qiu, Yftah Ziser, A. Korhonen, E. Ponti, and Shay B. Cohen. 2023. Detecting and mitigating hallucinations in multilingual summarisation. *ArXiv*, abs/2305.13632.

Krista Stinne Greve Rasmussen and Kirsten Vad. 2025. Kommentarer, registre, data. om kommentering af grundvigs værker i et digitalt paradigme. In Jeppe Barnwell, Simon Skovgaard Boeck, and Karen Skovgaard-Petersen, editors, *Ingen Kommentarer*, pages 63–81. Det Danske Sprog- og Litteraturselskab, Copenhagen.

Julian Risch, Marvin Schröder, Johannes Daxenberger, and Iryna Gurevych. 2022. Semantic answer similarity for evaluating question answering systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4784–4797.

Laurent Romary. 2012. Book review: Natural language processing for historical texts by michael piotrowski. *Computational Linguistics*, 40:231–233.

Matthias Schoffel, Marinus Wiedner, Esteban Garces Arias, Paula Ruppert, Christian Heumann, and M. Aßenmacher. 2025. Modern models, medieval texts: A pos tagging study of old occitan. *ArXiv*, abs/2503.07827.

Thomas Scialom, Paul-Alexis Dray, P. Gallinari, S. Lamprier, Benjamin Piwowarski, Jacopo Staiano, and Alex Wang. 2021. Questeval: Summarization asks for fact-based evaluation. pages 6594–6604.

Gemma Team. 2025a. Gemma 3n.

Qwen Team. 2025b. Qwen3 technical report.

Daniel Varab and Natalie Schluter. 2020. DaNewsroom: A large-scale Danish summarisation dataset. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6731–6739, Marseille, France. European Language Resources Association.

Lars S. Vikør. 2022. Rettskrivingsreform i store norske leksikon på snl.no. In *https://snl.no/rettskrivingsreform*.

Qingyue Wang, Liang Ding, Yanan Cao, Zhiliang Tian, Shi Wang, Dacheng Tao, and Li Guo. 2023. Recursively summarizing enables long-term dialogue memory in large language models. *ArXiv*, abs/2308.15022.

Zongda Wu, Li Lei, Guiling Li, Hui Huang, Chengren Zheng, Enhong Chen, and Guandong Xu. 2017. A topic modeling based approach to novel document automatic summarization. *Expert Systems with Applications*, 84:12–23.

Ming Zhang, Chengzhang Li, Meilin Wan, Xuejun Zhang, and Qingwei Zhao. 2023. Rouge-sem: Better evaluation of summarization using rouge combined with semantics. *Expert Syst. Appl.*, 237:121364.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Chao Zhao, Faeze Brahman, Kaiqiang Song, Wenlin Yao, Dian Yu, and Snigdha Chaturvedi. 2022. Narrasum: a large-scale dataset for abstractive narrative summarization. *arXiv preprint arXiv:2212.01476*.

Naitian Zhou, David Bamman, and Isaac L. Bleaman. 2025. Culture is not trivia: Sociocultural theory for cultural NLP. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25869–25886, Vienna, Austria. Association for Computational Linguistics.