Evaluating Open-Source LLMs for Text Summarization and Named Entity Recognition in Apartheid Witness Reports

Pauline Kister
Technical University of Munich
pauline.kister@tum.de

Miriam Schirmer
Northwestern University
miriam.schirmer@northwestern.edu

Abstract

This work investigates the extent to which open-source Large Language Models (LLMs) can improve accessibility of unstructured historical documents by performing abstractive summarization and fine-grained Named Entity Recognition (NER) for role classification and violation types. We evaluate open-source LLMs in zero-shot settings and apply these tasks to witness testimonies collected by the South African Truth and Reconciliation Commission (TRC), which archived a large body of text documenting human rights violations during apartheid. Despite their historical significance, these texts are difficult to access due to their length, lack of standardized structure, and the absence of systematic indexing. Open-source LLMs show strong performance in summarization, with most models surpassing non-LLM baselines (maximum BERTScore 0.77), while NER performance remains limited (maximum F1-score 0.61). Results suggest a trade-off in which stylistic fluency is prioritized over factual precision. A two-stage pipeline, summarization followed by NER on LLM summaries, leads to measurable improvements.

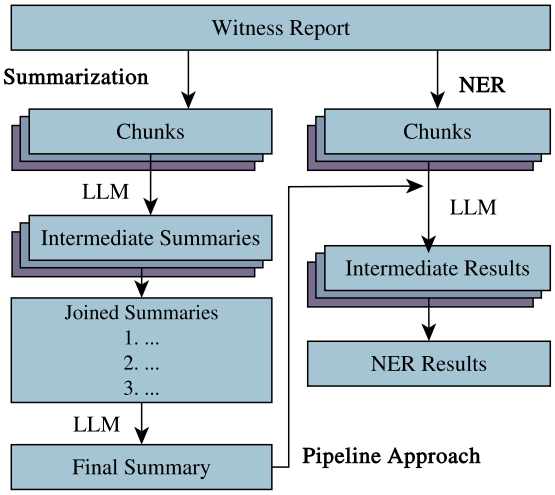

Figure 1: Study Overview.

1 Introduction

The preservation of historical records is essential for understanding the past and informing future generations. Natural Language Processing (NLP) tasks such as summarization and Named Entity Recognition (NER) can support this goal by making large, unstructured archives more accessible and searchable, particularly through transparent and reproducible open-source approaches.

In South Africa, the Truth and Reconciliation Commission (TRC)[1] produced an extensive collection of testimonies that document the experiences of victims and witnesses of human rights violations during apartheid. Prior research on witness testimonies from atrocity tribunals highlights the historical and legal value of such material (Keydar, 2020; Schirmer et al., 2022, 2025), suggesting that the TRC collection constitutes a valuable resource. However, it is difficult to access for a wider audience due to its volume, unstructured format, and the linguistic complexity of first-person testimonies. NLP can make such collections more accessible through automatic extraction and summarization of key information. Conventional NLP techniques frequently encounter difficulties with culture- and domain-specific vocabulary, irregular formatting, and historical contexts (Blouin et al., 2021). Although primarily designed for text generation, Large Language Models (LLMs) have demonstrated high performance on a wide range of NLP tasks, including information extraction (Hilgert et al., 2024; Kostina et al., 2025). Open-source LLMs of increasing quality can now be applied transparently and reproducibly, without task-specific training.

Building on this, we examine whether open-source LLMs can reliably extract key information

[1] https://www.justice.gov.za/trc/

Proceedings of the 6th International Conference on Natural Language Processing for the Digital Humanities, pages 390–410

from complex, unstructured texts, focusing on two tasks in particular: abstractive summarization and NER of victim and witness names, as well as the identification of violations discussed in the respective hearing. While the extraction tasks technically combine aspects of NER, classification, and information extraction, we refer to them as NER tasks throughout this work for simplicity. This is consistent with established work on fine-grained NER, where entity types extend beyond standard categories such as PERSON or LOCATION to include domain-specific roles (Choi et al., 2018; Zhang et al., 2020a).

LLMs are particularly promising for these tasks, since most existing NER models are limited to standard entity types and require task-specific fine-tuning. In contrast, our task demands the identification of fine-grained, context-dependent roles, such as distinguishing between victims and witnesses, without additional training. This paper's contributions are as follows:

1. We show that **open-source LLMs are highly effective at abstractive summarization** of long and complex historical documents (BERTScores of up to 0.77), being competitive with commercial models, while achieving moderate performance on entity detection (F1-scores up to 0.61).

2. We present performance trends with a **comparative analysis of different open-source models**;

3. introduce a **pipeline approach** that first summarizes the text and then applies NER, resulting in significantly improved results; and

4. provide a **curated dataset** of 200 witness reports corresponding to over 1,000 pages of court documents, including human- and LLM-generated summaries and structured metadata such as victim and witness names and the violations discussed.

We use free and open-source language models, with a focus on small systems, aiming to reduce computational costs and improve usability in research contexts with limited resources. All data and code are available online.[2]

[2] https://github.com/ge56dit/llm-apartheid-summarization-ner

2 LLMs for Summarization and NER

For abstractive summarization, LLMs have been established as the state-of-the-art method, consistently outperforming previous approaches. In contrast, LLMs are rarely the method of choice for NER due to their focus on text generation tasks. Despite the challenges for LLMs in NER, their flexibility and high context-understanding make them a promising option when dealing with fine-grained NER and classification on complex texts.

Summarization for Long Inputs. With the introduction of LLMs, the capabilities of abstractive summarization methods have improved significantly. Models such as GPT-3 (Brown et al., 2020) have produced coherent and factual summaries, being fine-tuned to fit a specific task (Alexandr et al., 2021; Rallapalli et al., 2025) or used in zero- or few-shot settings (Leiva-Araos et al., 2025; Liu and Healey, 2023; Pokale et al., 2023).

One of the biggest challenges in summarization is how to deal with long inputs. A document is usually considered long if it exceeds the maximal input length of state-of-the-art models, causing biases and lower performance (Ravaut et al., 2024; Hosseini et al., 2024). The most recent models allow for very long to theoretically unlimited inputs; however, hardware limitations still require strategies for long inputs. Recent work suggests solutions, such as chunking the text and merging the results, intelligently truncating parts of the text that are unlikely to influence downstream tasks, or preprocessing with extractive summarization methods (Chang et al., 2024; Hosseini et al., 2024).

Summarization for Historical Documents. When working with historical texts, language models often encounter problems such as unknown vocabulary or scanning errors (Lyu et al., 2021). Task-independent word embeddings trained on historical documents like HistBERT (Wenjun Qiu and Xu, 2022) can partially address these problems. In addition, several methods have been developed specifically to summarize historical texts, e.g., using temporal information to cluster sentences and produce extractive summaries (Gung and Kalita, 2012).

For in-context learning with LLMs specifically for historical documents, research is more scarce. Murugaraj et al. (2025) conclude that fine-tuned neural models still outperform general-purpose LLMs in this domain and that the historical source

of their input texts contributes to the lower performance of LLMs, suggesting that LLMs struggle with the challenges of historical texts. In contrast, Zhang et al. (Zhang et al., 2024a) report that GPT3.5 achieved better results than other models fine-tuned for summarization of historical texts, but is still struggling with hallucinations.

NER for Historical Documents. Just as for summarization, NER in historical texts poses distinct challenges due to archaic orthography, digitization errors, and the scarcity of annotated data. Historical texts often contain OCR-induced errors, which degrade NER performance (Hamdi et al., 2019). However, careful evaluation and targeted error-correction methods can substantially mitigate these effects (Boros et al., 2020; Trias et al., 2021; Won et al., 2018). NER systems for historical documents primarily use deep learning, often combined with pretrained transformer-based embeddings (Aguilar et al., 2017; Ehrmann et al., 2016, 2023; Lample et al., 2016; Won et al., 2018). The use of transformer-based models led to improved results, especially in domains where limited data is available, such as historical texts (Schweter and Baiter, 2019) or in low-resource languages such as those of South Africa (Hanslo, 2022).

While LLMs excel at tasks such as summarization, their performance for NER has been ambiguous (Wang et al., 2023). In the historical domain, the quality of results for NER is even lower due to errors and domain-specificity (Tudor et al., 2025; González-Gallardo et al., 2024; Zhang and Colavizza, 2025). Methods to enhance the performance of LLMs include improving few-shot learning (Ashok and Lipton, 2023; Jiang et al., 2024b), cascading multiple LLMs (Luo et al., 2025), or self-correcting methods (Polak and Morgan, 2024; Wang et al., 2023; Xiao et al., 2024).

A Pipeline Approach to Fine-Grained NER Classification in TRC Texts. While most NER systems focus on high-level categories (e.g., person, location), our approach requires fine-grained classification to identify victims and witnesses. Existing NER models like TexSmart (Zhang et al., 2020a) can distinguish between more fine-grained labels such as towns and countries or identify victims in crime reports (Choi et al., 2018; Schirmer et al., 2024a). However, most fine-grained systems lack domain-specific categories, and pretraining constraints make it difficult to add new labels without retraining.

To address these challenges, we employ a pipeline approach to improve the results by first performing summarization, then NER. While it is known that splitting a task into multiple steps can improve an LLM's ability to follow instructions (Xie et al., 2024; Khot et al., 2023), this specific approach has rarely been investigated in detail. Nevertheless, NER on summarized texts has been performed successfully in the medical domain (Sasikala et al., 2024). There is also evidence that incremental summarization can help smaller language models perform sentiment classification, another task that is usually unrelated to text generation (Ma et al., 2018).

3 Dataset Collection and Annotation

We compile a new dataset from publicly available documents from the Truth and Reconciliation Commission (TRC). The TRC was established in 1996, a few years after the end of apartheid in South Africa, to investigate politically motivated human rights violations, provide reparation and rehabilitation, and grant amnesty (Truth and Comission, 1998). It collected more than 1300 witness testimonies about human rights violations, of which 200 were used in this study. Witness testimonies provide a direct and less formal retelling of events. They are mostly unscripted, offering emotionally charged but very detailed narratives. To ensure witnesses could testify in the language of their choice, translators were provided for 12 different languages (Truth and Comission, 1998). Most transcripts are available in English, usually without an indication of whether they were translated. This does not reduce the dataset's quality, as the translations were produced by professional court translators and preserve not only meaning but also the style, including errors and choice of wording (Gilbert and Heydon, 2021). Similar corpora have been compiled from genocide-related court proceedings (Schirmer et al., 2023, 2024b).

In contrast to the training of a neural network specialized in witness reports and summarization or NER, the LLM approach has the advantage of not requiring any training data and therefore no annotations. However, to validate model performance, reliable reference data is necessary.

To extend and diversify the validation corpus, three trained experts annotated a subset of 200 documents (ranging from $\sim$ 2 to 110 pages). Experts were part of the research team, including one

postdoc, one master's student, and one research assistant. All of the annotators followed detailed annotation guidelines (see A.1). To ensure sufficient annotator agreement, all annotators first annotated 10 identical documents. Inter-annotator agreement was measured using task-appropriate metrics: BERTScore for summarization, where free-text outputs cannot be compared categorically, and F1-score for the extraction tasks. Since we use the same metrics for model evaluation, these values indicate how much of the observed error can be attributed to the subjectivity of the task instead of a failure of the model. The agreement reached a good BERTScore of 0.75 for summaries, a very good F1-score of 0.94 for personal names, and an F1-score of 0.62 for type of crime labeling. The crime category is often named implicitly in the witness reports, and category boundaries may overlap, which makes the task highly subjective and constrains the achievable agreement. Since human summarization is also known to be a subjective task (Liu et al., 2023), we consider these results to be acceptable. In a group meeting, we discussed differences in labeling and adapted the guidelines accordingly. Each annotator then continued to annotate randomly chosen documents. The final dataset used for this study contained 200 transcripts, including over 1,000 pages of witness recounts describing human rights violations during apartheid, collected by the TRC and recorded between 1996 and 1997.

4 Methods

The models we compared were TinyLlama (Touvron et al., 2023), Mistral (Jiang et al., 2023), Mixtral (Jiang et al., 2024a), OpenHermes (Teknium, 2023), Gemma3 (Team et al., 2025), and Phi-4 (Abdin et al., 2024). They are free to use and open-source, which ensures transparency and reproducibility. In addition, we included GPT-4o-mini (OpenAI, 2024) as a comparison to commercial models. To explore the influence of instruction fine-tuning on the model's responses, we also applied a second version of Mistral, uMistral, which receives unstructured prompts (default system prompt; instructions and input text in a single message) (see Table 1 for a model overview). Excluding the baseline models, we used the same models for both tasks. Hyperparameter tuning was conducted for all models with a manually supported grid search, mostly with parameters controlling the abstraction level (e.g., temperature, beam search parameters) and penalties for repetition and output length (A.2.4, Table 4).

For all prompted models, the prompt template was identical and systematically engineered (see A.2 for the final templates). Due to the limited context length of the models, reports sometimes had to be split into chunks at the sentence level before they could be passed to the model. To avoid excessive splitting and a loss of context, the prompt template had to be as short as possible, which is why we only consider zero-shot settings. Each text chunk was wrapped in this template and then formatted as expected by the instruction tuning of each model: One message for the system prompt, one for the instructions, and one for the input text. uMistral only received a single message consisting of the input text and instructions, disabling some aspects of the instruction tuning.

4.1 Summarization

This work focuses on abstractive summarization because LLMs are more suited to rephrase and generalize the original text than to determine the importance of individual sentences. We compared the performance of the LLMs against several smaller baseline models fine-tuned for summarization: BART (Lewis et al., 2019), T5 (Raffel et al., 2019), and Pegasus (Zhang et al., 2019). We chose general-purpose models rather than models specialized for historical texts, because available encoder-only models like HistBERT (Wenjun Qiu and Xu, 2022) are not comparable to fully trained encoder-decoder summarization models.

Experimental Setup. The models were prompted to write a short summary that describes the course of events in each text chunk. For the full prompt, see A.2.1. The resulting summaries were joined together. Experiments showed that clearly separating and enumerating the summaries is more effective than naive concatenation. The joined summaries were then passed back to the model to create a single, coherent text. If the joined summary was longer than the model's input, it had to be split again, and the process was repeated recursively.

We conducted additional experiments to assess how context window size and recursive summarization affect summary quality by uniformly limiting input length across models and by truncating intermediate summaries. These test how reduced con-

Table 1: Overview of the models.

Name	Version	Params	Context	Key Feature
T5	long-t5-tglobal-base	220M	4,096	Summarization-only, long context
BART	bart-large-cnn	403M	1,024	Summarization-only, BERT-based
Pegasus	pegasus-cnn-dailymail	568M	1,024	Summarization-only
TinyLlama	Chat-v1.0	1.1B	2,048	Highly compact LLM
Mistral	Instruct-v0.3	7.3B	32,768	Mistral with prompt formatting
uMistral	Instruct-v0.3	7.3B	32,768	Mistral without prompt formatting
Mixtral	Instruct-v0.1	8x7B	32,768	Mixture of experts
OpenHermes	2.5-Mistral-7B	7.3B	32,768	Strongly instruction-tuned Mistral
Gemma3	gemma-3-4b-it	4B	128,000	Good long-context abilities
Phi-4	Phi-4	14B	32,768	Good factuality and reasoning
GPT	GPT-4o-mini	unknown	128,000	Commercial model

text and recursive merging trade off against error propagation (see A.2.5).

Summarization Evaluation. The most straightforward way to evaluate a summary is to compare it to a human-written reference summary treated as ground truth. We rely on two widely adopted reference-based metrics: ROUGE (Lin, 2004) and BERTScore (Zhang et al., 2020b). **ROUGE** is the average of three different metrics: the overlap of unigrams between the summary and reference, the overlap of bigrams, and the length of the longest common subsequence. **BERTScore** is the cosine similarity between the BERT embeddings of the model output and the reference. It compares texts on a semantic level and is less sensitive to paraphrasing than ROUGE.

Since reference summaries are not always available, we additionally use reference-free metrics. The **Unique Bigram Ratio** measures lexical diversity, the **Summarization Ratio** captures compression rate, and **Semantic Similarity** computes the BERT-based similarity between generated text and the source text to assess faithfulness, with high values potentially indicating extractiveness.

4.2 Named Entity Recognition

In the case of the TRC dataset, an important piece of information to extract is personal names, as they enable searching for documents related to a specific case. Similarly important is information about the type of crime that was committed (e.g., "shooting", "torture").

Experimental Setup. The NER process was similar to the process for summarization. The original text was split into chunks of maximum size for the model's context window, taking into account sentence boundaries. For personal names, models were prompted to extract the names of victims and witnesses from the text, while receiving definitions of who is considered to be a victim or witness (see the full prompt at A.2.2). Important improvements during prompt engineering included the phrase "most important names" to reduce false positives, and the instruction of what to do when the victim's or witness's names are not mentioned to avoid hallucinations. For violation classification, the model received short definitions for possible violation classes and was prompted to cite one or more of these labels.

Since conventional NER models cannot distinguish between different crime types and roles, they were unable to solve the exact task we propose. We still include a SpaCy (Honnibal et al., 2020) baseline for the victim and witness name recognition task to serve as a reference point for general-purpose person name recognition rather than role-specific entity extraction.

We conducted additional experiments to control for the effects of entity combination strategies and input length by comparing fuzzy matching and LLM-based aggregation and by reducing the maximal input to 1,024 tokens for all models (see A.2.6). In addition, we improved the results with a pipeline approach, where we applied NER on the reference summaries as well as on the model-generated summaries to assess whether summarization simplifies NER for long TRC documents.

NER Evaluation. For extracting personal names, the model was instructed to answer in JSON format. The first JSON object in the text was considered the model's answer; any text around it was ignored. For

the crime classification, the model was instructed to cite keywords from a given list (see the prompt A.2.3 for a definitive list). Only classes from the list were considered in the evaluation. We extracted lists of victims, witnesses, and classes from the model's answers and calculated precision, recall, and the micro F1-score. The token set ratio allowed for soft matching for personal names.

We additionally conducted a qualitative manual error analysis for summarization and NER in parallel. Starting with a subset of the evaluated documents, we compared the model outputs to both the reference annotations and the original source texts. This allowed the identification of potential explanations for model failures and document-specific challenges. Emerging patterns were then validated by searching for more instances in which the model showed similar behavior. In parallel, we identified outliers in the evaluation metrics, such as unusually long summaries or instances with many false positives. These cases were examined manually to determine the cause of the errors. Through this process, several patterns were identified that can give insights into the shortcomings of the models and the difficulties of the dataset.

5 Results

For summarization, GPT, Mixtral, and both Mistral models perform best with BERTScores of around 0.7 to 0.8. In contrast, the LLMs' NER performance is moderate. GPT, uMistral, and Mixtral are the only models that reach an F1-score close to 0.6. Most models yield results with high recall, but low precision. The pipeline approach improves these results.

5.1 Summarization

Across all models, performance varied notably in both ROUGE and BERTScore evaluations (see Tables 2 and A.3.1, Figure 4 and 5). In ROUGE, the Mistral models achieved the highest average score (both 0.26), followed closely by Mixtral and GPT (0.25). OpenHermes and Gemma3 still surpass the best baseline model, BART, which reached a ROUGE score of 0.19. Lower results were obtained by TinyLlama (0.17), Pegasus (0.14), and T5 (0.12). BERTScore rankings followed a similar trend. GPT led with a score of 0.78, closely followed by Mistral (0.77), uMistral (0.77), and Mixtral (0.75). BART (0.64) scored in the mid range, surpassing Phi-4 (0.63) and OpenHermes(0.64), while the other baseline models and TinyLlama scored significantly lower.

The unique bigram ratio indicates that BART, OpenHermes, and GPT models have greater lexical diversity and less repetition. The low score of T5 (0.64) hints at much repetition in the summaries. The variations in the summarization ratio indicate that models with lower scores, such as BART and Pegasus, may not have been able to cover all important facts. In comparison, models with a higher score, like Phi4 and GPT, may include too much unnecessary detail, although a high score does not say much about the factuality. Models like uMistral and GPT, which have a high similarity to the source text in combination with a good BERTScore, produce highly factual summaries. T5's high similarity and low BERTScore point towards highly extractive summaries that directly cite the original text.

Overall, LLMs outperformed other architectures across surface-level and semantic metrics. The exception is TinyLlama, which does not surpass the best baseline with either score. The highest-ranking non-LLM model (BART) is outperformed by the highest-ranking LLM (GPT) by 0.13 in BERTScore, and the highest-ranking open-source model (uMistral) by 0.11.

Most models were robust to input and output length constraints, with only minor BERTScore changes, including slight drops for most models and improved results for OpenHermes (A.3.1, Figure 6).

5.2 Named Entity Recognition

NER results revealed substantial differences between the models (Table 3; A.3.2, Figure 7). uMistral achieved the best results (F1-score 0.60), followed by GPT (0.59), Mixtral (0.58), and Mistral (0.51). All of these models achieved a much higher recall than precision, indicating that the models included false names in the list. Gemma3, Phi-4, and OpenHermes scored lower, but with a better precision and recall balance.

The SpaCy baseline could detect personal names in the text, but was by design not able to distinguish between victims and witnesses or to classify the crime type. By extracting all names found in the text, it reached an F1-score of 0.19 in the victim and witness name detection. This supports our findings that non-LLM models struggle to solve the NER-based role classification task.

The better-performing models reached higher

Table 2: Results of the main summarization experiment. Mean and standard deviation over all witness reports. Values closest to the reference are marked in blue.

Model	BERTScore	ROUGE	Bigram ratio	Sum. ratio	Sim. to source
BART	0.6567± 0.131	0.1878± 0.086	0.9436± 0.045	0.0241± 0.018	0.4686± 0.086
T5	0.4186± 0.129	0.1219± 0.074	0.6492± 0.299	0.0938± 0.067	0.6210± 0.158
Pegasus	0.5574± 0.171	0.1386± 0.064	0.8581± 0.128	0.0162± 0.014	0.4427± 0.098
TinyLlama	0.5218± 0.114	0.1651± 0.096	0.8596± 0.044	0.1109± 0.098	0.5506± 0.091
Mistral	0.7708± 0.084	0.2621± 0.119	0.8268± 0.066	0.0987± 0.058	0.6386± 0.088
uMistral	0.7673± 0.077	0.2640± 0.115	0.8074± 0.078	0.1045± 0.061	0.6439± 0.085
Mixtral	0.7504± 0.153	0.2503± 0.115	0.8379± 0.071	0.0934± 0.069	0.6001± 0.142
OpenHermes	0.6374± 0.219	0.2068± 0.106	0.9130± 0.080	0.0623± 0.043	0.5435± 0.115
Gemma3	0.7292± 0.092	0.2197± 0.110	0.7993± 0.086	0.1335± 0.064	0.6819± 0.078
Phi-4	0.6304± 0.143	0.1812± 0.097	0.8192± 0.063	0.1672± 0.106	0.6688± 0.092
GPT-4	0.7832± 0.061	0.2480± 0.121	0.8832± 0.042	0.1124± 0.065	0.6483± 0.074
Reference	1.0	1.0	0.8946± 0.056	0.0556± 0.033	0.5348± 0.077

Table 3: NER results across models with precision, recall, and micro F1-score (main experiment). Best scores are marked in blue.

Model	Precision	Recall	F1-score
TinyLlama	0.2097	0.0132	0.0248
Mistral	0.4256	0.6316	0.5086
uMistral	0.5137	0.7399	0.6064
Mixtral	0.4906	0.7136	0.5814
OpenHermes	0.3828	0.1933	0.2569
Gemma3	0.3713	0.4028	0.3864
Phi-4	0.4778	0.2945	0.3644
GPT-4o-mini	0.5057	0.7126	0.5916

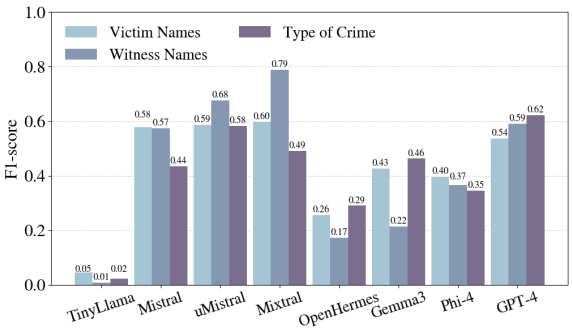

Figure 2: NER micro f1-scores for victim names, witness names, and violations.

scores at predicting witness names than victim names, indicating that this task is less complex (Figure 2). For the witness names, the highest F1-score of 0.79 is achieved by Mixtral, the model with the highest difference between victim and witness names. In contrast, the lower-performing models like OpenHermes and Gemma3 reached better results for the victim name. The crime classification scores are generally lower, with Gemma3 achieving surprisingly good results in contrast to a lower-performing Mistral.

Allowing the LLM to combine the results for each chunk independently through a final model call, thereby avoiding the possible errors introduced by fuzzy matching, results in slight improvement for the better-performing models, while models like OpenHermes and Gemma3 struggled with the combination (A.3.2, Figure 8).

5.3 The Pipeline Approach

Applying the **pipeline approach** that uses summaries as input for NER instead of the source text improved the results for 5 out of 8 models, with no changes for the other 3 models (see Figure 3). The results generally showed significant improvement over the main experiment results for the lower-performing models, and modest improvements for the higher-performing models, with a maximum improvement of 0.23 in F1-score for OpenHermes. NER on the human-written reference summary increased the F1-score for all models, especially the lower-performing models, sometimes by up to 0.4 in F1-score. These results represent an upper ceiling for the potential improvement that the pipeline approach could achieve on good summaries.

6 Discussion

The goal of this study was to evaluate open-source LLM performance for summarization and NER. The contrasting results – good scores for summarization and moderate scores for NER – show that, despite the LLMs' ability to solve a variety of

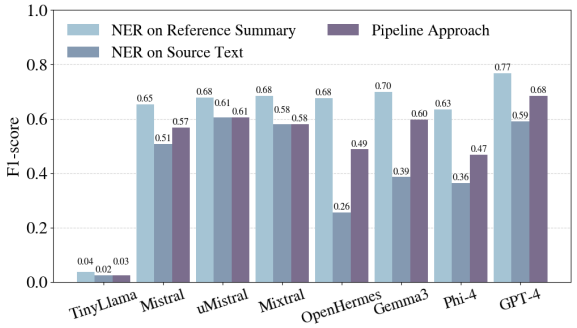

Figure 3: NER F1-scores on the reference summary, on the source text, and on the model-generated summary (pipeline approach).

tasks quickly without fine-tuning, they are mainly reliable for text generation tasks. Their greatest strength, their ability to produce coherent texts, is essential for summarization, but of less benefit for NER.

6.1 Summarization

Compared to NER and to the non-LLM models, summarization with LLMs yielded moderate to strong results, with the LLMs outperforming earlier neural approaches specifically fine-tuned for summarization. The difference between LLM and non-LLM models demonstrates that LLMs are a better choice for this task. The scale of their training data provides them with broader world knowledge and superior paraphrasing capabilities, a clear advantage, especially when considering that, due to their specificity and text format, the TRC reports require a high amount of background knowledge and abstraction.

Error Analysis. In general, the baseline models seem to struggle with overly extractive summaries, while the LLMs more often abstract too much, thereby generalizing to the point where the meaning is lost. Across models, distinct qualitative patterns emerge: TinyLlama sometimes replaced the content almost entirely with generic, hallucinated legal conclusions, while larger LLMs such as Mistral and especially Mixtral generated coherent and largely accurate narratives that recovered key details, but still occasionally included irrelevant or overly generic statements.

The following example is a summary excerpt by Mistral relating to the case of Peter Mabilo, a young man who was shot by the police (see A.4 for the full reference summary and more model summaries). Red coloring indicates hallucinations, green highlights relevant factual information, and blue text is not relevant for this summary:

> [...] Ms Mabozo, Luthli's grandmother, provided additional details about the shooting itself. She stated that after the police took Luthli's father away, they were told to take the children outside. Luthli was inside the house when the police shot him. The family found the house damaged and full of bullet holes [...]. In conclusion, the Commission expressed its sympathy for the family and vowed to investigate the circumstances surrounding Luthli's death [...].
> – Mistral, BERTScore 0.71, ROUGE 0.25

The last, irrelevant paragraph can be seen as an example of a generic statement typical for highly instruction-tuned models. Most models also struggle with reconstructing the course of events in temporal order, a known problem in NLP (Qiu et al., 2023; Barale et al., 2025)

6.2 Named Entity Recognition

NER results ranged from poor to moderate. Again, GPT, uMistral, and Mixtral achieved the best results. The quality of the NER results is hard to judge, since there is no existing baseline for this very specific task. The task differs from standard NER benchmarks in that it requires contextual role assignment rather than entity type detection. Other works that apply NER with LLMs can achieve better results with shorter texts, few-shot settings, and more extensive prompt engineering (Isaradech et al., 2024); however, it is also widely reported that LLMs struggle with NER (Lu et al., 2025). The results shown here are therefore to be expected when considering the difficulty of the task.

Interestingly, the highly instruction-aligned models performed worse than more general models, such as Mistral and Mixtral. While this may seem counterintuitive, those networks may be too specialized for such an unusual task. This effect can also be observed in the significant difference in performance with and without a correctly structured prompt for Mistral: The unstructured input and default system prompt partially disable the chat fine-tuning, resulting in outputs that are clearer and more concise. Therefore, uMistral, though being a relatively small open-source model, performs equally as good in NER as a commercial model like GPT-4o-mini.

Most models show higher recall than precision. This is also a result of the model's chat fine-tuning, which encourages it to provide very verbose and detailed answers (Zhang et al., 2024b). The models might therefore add false names to the result to appear more elaborate.

The mediocre performance of TinyLlama is likely due to its small size compared to the other networks. However, Phi-4, as the largest model, is outperformed by three different smaller models. This shows that as long as the models are sufficiently large to have decent language-understanding capabilities, the instruction training and specialization of the model seem to matter more than the number of parameters.

The success of the pipeline approach shows the effectiveness of splitting a task into multiple subtasks of less complexity. The summarization process cuts out parts of the text that are purely procedural and shifts the perspective and tone, which supports the entity recognition. Shorter inputs with fewer names simplify the victim and witness detection, while a direct and concrete retelling of the events of the crime helps with crime type classification. Observed patterns indicate that summarization not only shortens the input, but also changes the reports into a representation that is easier for the models to process.

Error Analysis. Typical mistakes made by most models mainly centered around challenges with naming conventions and the distinction between witnesses and victims. A recurring difficulty for all models was handling the multiple names by which a single person could be referred. For example, witnesses often used a victim's first name, while the commission referred to the same person by their family name. These connections were usually clear to human readers through context or by an early mention of the full name.

Cultural naming conventions added further complexity. Many people in the dataset followed the South African tradition of having two first names. The so-called home name often comes from Zulu or Xhosa tradition and is used by the close family, while the school name is used by the public (Suzman, 1994; Neethling, 2008). This led to false positives, when the model correctly classified both names as victims, but did not recognize that they belonged to the same person. For example, the name "Kwinda Daphne Tshinane" would be listed both as "Kwinda Tshinane" and "Daphne" in the model output. However, if both names were mentioned at least once in direct succession, the models would usually avoid this error.

For the NER task, hallucinations were less common: Many false positives were not full hallucinations, but actual names that appeared in the reports, such as family members or commissioners. This suggests that the models were generally able to identify names correctly, but struggled to classify them into the correct category.

7 Conclusion

This study showed that LLMs outperformed other models in summarization and were able to produce at least decent results in a fine-grained NER task that non-LLM models are not able to solve without further fine-tuning. The best performing open-source LLMs were competitive with GPT. However, the optimization of LLMs towards coherence and instruction following comes at the expense of accuracy. For summarization, this leads to overly long outputs with hallucinations and generic filler sentences. In NER, this leads models to avoid empty outputs and to label almost any detected name as either a victim or a witness, resulting in high recall but low precision.

Unlike general-purpose NLP tasks, mistakes in this context can have ethical and historical consequences. Omitting a victim's name or falsely attributing responsibility for a crime can alter the record and lead to wrongful accusations. LLM summaries may also be easier to read and more coherent than the witness reports, which might mislead readers due to higher credibility despite potential hallucinations. While LLMs offer new opportunities to make historical texts more accessible and easier to analyze, at the current state, the application of LLMs to sensitive historical datasets should always be accompanied by careful human oversight and transparency about model limitations. Pipeline approaches with models specialized in the task, combined with the flexibility of LLMs, may improve the results in future work.

8 Limitations

One essential limitation of these results, particularly for summarization, is the observation that numerical evaluation methods are not entirely reliable when assessing the quality of the results. While BERTScore and ROUGE score capture summary quality relatively well, they do not always align

with human judgment and are dependent on the quality of the reference summary. Even embedding-based BERTScore occasionally fails when it comes to more complex semantics (Zhang et al., 2020b; Fabbri et al., 2021). We therefore interpret the results primarily in terms of overall trends rather than absolute values and employ multiple evaluation metrics to mitigate metric-specific biases. Reference-free metrics, too, cannot reliably indicate the quality of a summary. However, when observed in combination with reference-based metrics, they can give insight into the model's strengths and weaknesses, such as the repetitiveness of T5 or the verbosity of Mixtral.

Although the ROUGE scores reported here may seem low, they are relatively good results considering the task. Paraphrasing is part of the process of abstractive summarization, and the wording of two summaries will inevitably differ. ROUGE-scores for this task are usually in a similar range to the results achieved here (Fabbri et al., 2021; Lam et al., 2022).

The raw witness transcripts published by the TRC were initially recorded on tape and later transcribed without modification. As a result, they preserve all the imperfections of natural speech, including stuttering, repetition, unfinished sentences, filler words, and word mix-ups. The emotional nature of the testimonies often intensified this. Additional errors arise from the recording process itself, including microphone malfunctions and losses of several seconds to minutes when tapes were flipped. The transcription process introduced additional errors, such as missing speaker labels, misspelled names, and arbitrary line breaks.

The cultural and historical background of the TRC-dataset was certainly a challenge for most models. Understanding these reports requires not only good reasoning abilities but also detailed knowledge about apartheid and South African history, which the models seem to lack. LLMs are known to show a bias against non-Western cultures (Tao et al., 2024); it is therefore likely that an underrepresentation of South African historical and cultural material in the training corpora of LLMs has contributed to the models' shortcomings. If LLMs are used in practice for tasks on cultural datasets, special training methods are needed to mitigate this bias (Gallegos et al., 2024; Li et al., 2024).

The results presented here open up possibilities for future studies and improvements. Firstly, establishing a strong non-LLM baseline on the TRC dataset would allow clearer benchmarking. This would require more human-annotated data and the training of specialized neural models. Secondly, given that further instruction tuning did not improve NER performance, it may be worthwhile to evaluate non-instruction-tuned LLMs. Another possible improvement could be achieved with more complex pipeline approaches, where LLMs identify potential entities and a smaller classifier assigns roles. Alternative formulations of the NER task, such as tagging within the text rather than direct extraction, may also prove fruitful, although this would require networks with very large context windows. Thirdly, expanding experiments to other datasets could help disentangle whether the main challenges stem from the dialogue format, South African cultural references, or text length. The use of different yet similar annotated datasets may also enable fine-tuned LLMs to learn domain-specific language.

9 Ethical Considerations

This research only relies on information that is publicly available on the TRC website. Witnesses' names and personal information are not disclosed beyond the information that has been published by the TRC. Testimonies from mass atrocity proceedings carry particular psychological weight for those who engage with them (Schirmer, 2024). Human annotators were informed of and aware of the potentially violent content before the annotation process, with the ability to decline annotation at any time. They were given the chance to discuss any distressing material encountered during annotation and provided with a guide designed to aid in identifying changes in cognition and minimizing emotional risks associated with the annotation process (Kennedy et al., 2022).

Beyond annotation, ethical considerations should also address the potential impact of hallucinations. Here, a fabricated name in a victim or witness list could wrongfully tie falsely link someone to a human rights violation, or a could erase a victim from the record. LLM-generated summaries should therefore always be read alongside the original testimony, not instead of it.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael

Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, and 8 others. 2024. Phi-4 technical report. *Preprint*, arXiv:2412.08905.

Gustavo Aguilar, Suraj Maharjan, Adrian Pastor López-Monroy, and Thamar Solorio. 2017. A multi-task approach for named entity recognition in social media data. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 148–153, Copenhagen, Denmark. Association for Computational Linguistics.

Nikolich Alexandr, Osliakova Irina, Kudinova Tatyana, Kappusheva Inessa, and Puchkova Arina. 2021. Fine-tuning gpt-3 for russian text summarization. In *Data Science and Intelligent Systems*, pages 748–757, Cham. Springer International Publishing.

Dhananjay Ashok and Zachary C. Lipton. 2023. Promptner: Prompting for named entity recognition. *Preprint*, arXiv:2305.15444.

Claire Barale, Leslie Barrett, Vikram Sunil Bajaj, and Michael Rovatsos. 2025. Lextime: A benchmark for temporal ordering of legal events. *Preprint*, arXiv:2506.04041.

Baptiste Blouin, Benoit Favre, Jeremy Auguste, and Christian Henriot. 2021. Transferring modern named entity recognition to the historical domain: How to take the step? In *Proceedings of the Workshop on Natural Language Processing for Digital Humanities*, pages 152–162, NIT Silchar, India. NLP Association of India (NLPAI).

Emanuela Boros, Ahmed Hamdi, Elvys Linhares Pontes, Luis Adrián Cabrera-Diego, Jose G. Moreno, Nicolas Sidere, and Antoine Doucet. 2020. Alleviating digitization errors in named entity recognition for historical documents. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 431–441, Online. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Yapei Chang, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024. Booookscore: A systematic exploration of book-length summarization in the era of llms. *Preprint*, arXiv:2310.00785.

Eunsol Choi, Omer Levy, Yejin Choi, and Luke Zettlemoyer. 2018. Ultra-fine entity typing. *Preprint*, arXiv:1807.04905.

Maud Ehrmann, Giovanni Colavizza, Yannick Rochat, and Frédéric Kaplan. 2016. Diachronic evaluation of ner systems on old newspapers. page 97–107, Bochum, Germany. Bochumer Linguistische Arbeitsberichte.

Maud Ehrmann, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. 2023. Named entity recognition and classification in historical documents: A survey. *ACM Computing Surveys*, 56(2):1–47.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Tong Yu, Hanieh Deilamsalehy, Ruiyi Zhang, Sungchul Kim, and Franck Dernoncourt. 2024. Self-debiasing large language models: Zero-shot recognition and reduction of stereotypes. *Preprint*, arXiv:2402.01981.

David Gilbert and Georgina Heydon. 2021. Translated transcripts from covert recordings used for evidence in court: Issues of reliability. *Frontiers in Communication*, 6:779227.

Carlos-Emiliano González-Gallardo, Hanh Thi Hong Tran, Ahmed Hamdi, and Antoine Doucet. 2024. Leveraging open large language models for historical named entity recognition. In *Linking Theory and Practice of Digital Libraries*, pages 379–395, Cham. Springer Nature Switzerland.

James Gung and Jugal Kalita. 2012. Summarization of historical articles using temporal event clustering. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 631–635.

Ahmed Hamdi, Axel Jean-Caurant, Nicolas Sidere, Mickaël Coustaty, and Antoine Doucet. 2019. An analysis of the performance of named entity recognition over ocred documents. In *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 333–334.

Ridewaan Hanslo. 2022. Deep learning transformer architecture for named entity recognition on low resourced languages: State of the art results. In *Proceedings of the 17th Conference on Computer Science and Intelligence Systems*, volume 30 of *FedCSIS 2022*, page 53–60. IEEE.

Lukas Hilgert, Danni Liu, and Jan Niehues. 2024. Evaluating and training long-context large language models for question answering on scientific papers. In *Proceedings of the 1st Workshop on Customizable NLP: Progress and Challenges in Customizing NLP for a Domain, Application, Group, or Individual (CustomNLP4U)*, pages 220–236, Miami, Florida, USA. Association for Computational Linguistics.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spacy: Industrial-strength natural language processing in python.

Peyman Hosseini, Ignacio Castro, Iacopo Ghinassi, and Matthew Purver. 2024. Efficient solutions for an intriguing failure of llms: Long context window does not mean llms can analyze long sequences flawlessly. *Preprint*, arXiv:2408.01866.

Natthanaphop Isaradech, Andrea Riedel, Wachiranun Sirikul, Markus Kreuzthaler, and Stefan Schulz. 2024. Zero-and few-shot named entity recognition and text expansion in medication prescriptions using chatgpt. *arXiv preprint arXiv:2409.17683*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, and 7 others. 2024a. Mixtral of experts. *Preprint*, arXiv:2401.04088.

Guochao Jiang, Zepeng Ding, Yuchen Shi, and Deqing Yang. 2024b. P-icl: Point in-context learning for named entity recognition with large language models. *Preprint*, arXiv:2405.04960.

Brendan Kennedy, Mohammad Atari, Aida Mostafazadeh Davani, Leigh Yeh, Ali Omrani, Yehsong Kim, Kris Coombs Jr, Shreya Havaldar, Gwenyth Portillo-Wightman, Elaine Gonzalez, and 1 others. 2022. Introducing the gab hate corpus: defining and applying hate-based rhetoric to social media posts at scale. *Language Resources and Evaluation*, 56(1):79–108.

Renana Keydar. 2020. Listening from afar: An algorithmic analysis of testimonies from the international criminal courts. *University of Illinois Journal of Law, Technology & Policy*, pages 55–84.

Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. 2023. Decomposed prompting: A modular approach for solving complex tasks. *Preprint*, arXiv:2210.02406.

Arina Kostina, Marios D. Dikaiakos, Dimosthenis Stefanidis, and George Pallis. 2025. Large language models for text classification: Case study and comprehensive review. *Preprint*, arXiv:2501.08457.

Khang Lam, Tuong Do, Nguyet-Hue Pham, and Jugal Kalita. 2022. *Vietnamese Text Summarization Based on Neural Network Models*, pages 85–96.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California. Association for Computational Linguistics.

Andrés Leiva-Araos, Bady Gana, Héctor Allende-Cid, José García, and Manob Jyoti Saikia. 2025. Large scale summarization using ensemble prompts and in context learning approaches. *Scientific Reports*, 15(1):10259.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461.

Cheng Li, Damien Teney, Linyi Yang, Qingsong Wen, Xing Xie, and Jindong Wang. 2024. Park: Boosting cross-cultural understanding in large language models. *Preprint*, arXiv:2405.15145.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Sengjie Liu and Christopher G. Healey. 2023. Abstractive summarization of large document collections using gpt. *Preprint*, arXiv:2310.05690.

Yixin Liu, Alexander R. Fabbri, Pengfei Liu, Yilun Zhao, Linyong Nan, Ruilin Han, Simeng Han, Shafiq Joty, Chien-Sheng Wu, Caiming Xiong, and Dragomir Radev. 2023. Revisiting the gold standard: Grounding summarization evaluation with robust human evaluation. *Preprint*, arXiv:2212.07981.

Qiuhao Lu, Rui Li, Andrew Wen, Jinlian Wang, Liwei Wang, and Hongfang Liu. 2025. Large language models struggle in Token-Level clinical named entity recognition. *AMIA Annu Symp Proc*, 2024:748–757.

Hanjun Luo, Yingbin Jin, Xinfeng Li, Xuecheng Liu, Ruizhe Chen, Tong Shang, Kun Wang, Qingsong Wen, and Zuozhu Liu. 2025. Dynamicner: A dynamic, multilingual, and fine-grained dataset for llm-based named entity recognition. *Preprint*, arXiv:2409.11022.

Lijun Lyu, Maria Koutraki, Martin Krickl, and Besnik Fetahu. 2021. Neural OCR post-hoc correction of historical corpora. *Transactions of the Association for Computational Linguistics*, 9:479–493.

Shuming Ma, Xu Sun, Junyang Lin, and Xuancheng Ren. 2018. A hierarchical end-to-end model for jointly improving text summarization and sentiment classification. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4251–4257. International Joint Conferences on Artificial Intelligence Organization.

Keerthana Murugaraj, Salima Lamsiyah, and Christoph Schommer. 2025. Abstractive summarization of historical documents: A new dataset and novel method using a domain-specific pretrained model. *IEEE Access*, 13:10918–10932.

Bertie Neethling. 2008. Xhosa first names: A dual identity in harmony or in conflict? *Names: A Journal of Onomastics*, 56(1):32–38.

OpenAI. 2024. Gpt-4o technical report. *arXiv preprint arXiv:2405.19102*.

Sairaj Pokale, Karan Taware, Gavin Fernandes, Sakshi Kangane, Parth Bhosale, and Laxmi Bewoor. 2023. Text summarization: Gpt perspective. In *2023 3rd Asian Conference on Innovation in Technology (ASIANCON)*, pages 1–7.

Maciej P. Polak and Dane Morgan. 2024. Extracting accurate materials data from research papers with conversational language models and prompt engineering. *Nature Communications*, 15(1):1569.

Yifu Qiu, Zheng Zhao, Yftah Ziser, Anna Korhonen, Edoardo M. Ponti, and Shay B. Cohen. 2023. Are large language models temporally grounded? *Preprint*, arXiv:2311.08398.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683.

Swati Rallapalli, Shannon Gallagher, Andrew O. Mellinger, Jasmine Ratchford, Anusha Sinha, Tyler Brooks, William R. Nichols, Nick Winski, and Bryan Brown. 2025. Fine-tuning llms for report summarization: Analysis on supervised and unsupervised data. *Preprint*, arXiv:2503.10676.

Mathieu Ravaut, Aixin Sun, Nancy F. Chen, and Shafiq Joty. 2024. On context utilization in summarization with large language models. *Preprint*, arXiv:2310.10570.

D Sasikala, R Sudarshan, and S. Sivasathya. 2024. Harnessing llms for medical insights:ner extraction from summarized medical text. In *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–6.

Miriam Schirmer, Christian Brechenmacher, Endrit Jashari, and Jürgen Pfeffer. 2024a. Gentrac: A tool for tracing trauma in genocide and mass atrocity court transcripts. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7666–7671.

Miriam Schirmer, Udo Kruschwitz, and Gregor Donabauer. 2022. A new dataset for topic-based paragraph classification in genocide-related court transcripts. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4504–4512.

Miriam Schirmer, Tobias Leemann, Gjergji Kasneci, Jürgen Pfeffer, and David Jurgens. 2024b. The language of trauma: Modeling traumatic event descriptions across domains with explainable ai. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13224–13242.

Miriam Schirmer, Isaac Misael Olguín Nolasco, Edoardo Mosca, Shanshan Xu, and Jürgen Pfeffer. 2023. Uncovering trauma in genocide tribunals: An nlp approach using the genocide transcript corpus. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, pages 257–266.

Miriam Schirmer, Jürgen Pfeffer, and Sven Hilbert. 2025. Talking about torture: A novel approach to the mixed methods analysis of genocide-related witness statements in the khmer rouge tribunal. *Journal of Mixed Methods Research*, 19(1):83–102.

Miriam Katharina Doris Schirmer. 2024. *Natural language processing for violence studies: Investigating trauma and online aggression*. Ph.D. thesis, Technische Universität München.

Stefan Schweter and Johannes Baiter. 2019. Towards robust named entity recognition for historic German. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 96–103, Florence, Italy. Association for Computational Linguistics.

Susan M. Suzman. 1994. Names as pointers: Zulu personal naming practices. *Language in Society*, 23(2):253–272.

Yan Tao, Olga Viberg, Ryan S Baker, and René F Kizilcec. 2024. Cultural bias and cultural alignment of large language models. *PNAS Nexus*, 3(9):pgae346.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Teknium. 2023. Openhermes 2.5: An open dataset of synthetic data for generalist llm assistants.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Fernando Trias, Hongming Wang, Sylvain Jaume, and Stratos Idreos. 2021. Named entity recognition in historic legal text: A transformer and state machine ensemble method. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 172–179, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Truth and Reconciliation Comission. 1998. Truth and reconciliation commission of south africa report.

Crina Tudor, Beata Megyesi, and Robert Östling. 2025. Prompting the past: Exploring zero-shot learning for named entity recognition in historical texts using prompt-answering LLMs. In *Proceedings of the 9th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2025)*, pages 216–226, Albuquerque, New Mexico. Association for Computational Linguistics.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang. 2023. Gpt-ner: Named entity recognition via large language models. *Preprint*, arXiv:2304.10428.

Wenjun Qiu and Yang Xu. 2022. Histbert: A pre-trained language model for diachronic lexical semantic analysis.

Miguel Won, Patricia Murrieta-Flores, and Bruno Martins. 2018. Ensemble named entity recognition (ner): Evaluating ner tools in the identification of place names in historical corpora. *Frontiers in Digital Humanities*, Volume 5 - 2018.

Le Xiao, Yunfei Xu, and Jing Zhao. 2024. Llm-der:a named entity recognition method based on large language models for chinese coal chemical domain. *Preprint*, arXiv:2409.10077.

Yuanzhen Xie, Xinzhou Jin, Tao Xie, Matrixmxlin Matrixmxlin, Liang Chen, Chenyun Yu, Cheng Lei, Chengxiang Zhuo, Bo Hu, and Zang Li. 2024. Decomposition for enhancing attention: Improving LLM-based text-to-SQL through workflow paradigm. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10796–10816, Bangkok, Thailand. Association for Computational Linguistics.

Haisong Zhang, Lemao Liu, Haiyun Jiang, Yangming Li, Enbo Zhao, Kun Xu, Linfeng Song, Suncong Zheng, Botong Zhou, Jianchen Zhu, Xiao Feng, Tao Chen, Tao Yang, Dong Yu, Feng Zhang, Zhanhui Kang, and Shuming Shi. 2020a. Texsmart: A text understanding system for fine-grained ner and enhanced semantic analysis. *Preprint*, arXiv:2012.15639.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2019. PEGASUS: pre-training with extracted gap-sentences for abstractive summarization. *CoRR*, abs/1912.08777.

Ran Zhang, Jihed Ouni, and Steffen Eger. 2024a. Cross-lingual cross-temporal summarization: Dataset, models, evaluation. *Preprint*, arXiv:2306.12916.

Shibingfeng Zhang and Giovanni Colavizza. 2025. Named entity recognition of historical text via large language model. *Preprint*, arXiv:2508.18090.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020b. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

Yusen Zhang, Sarkar Snigdha Sarathi Das, and Rui Zhang. 2024b. Verbosity $\neq$ veracity: Demystify verbosity compensation behavior of large language models. *Preprint*, arXiv:2411.07858.

A Appendix

A.1 Annotation

Expert annotators followed the following instructions (summarized, the original guidelines can be found in this paper's code repository):

Summaries should be short, contain only factual information, and cover all important facts that are mentioned. The focus lies on the course of events surrounding the human rights violation, rather than on procedural matters or possible outcomes of the hearing. It is assumed that the report is accurate and complete. The purpose of the summary is to make the reports easier to search and filter, for example, for a specific type of crime or situation. Therefore, key information, such as the type of crime, names of the victims and witnesses, place, and time, must be mentioned. Additionally, the TRC typically considers information about the victim's political activities, the perpetrator's name and motive, and the long-term consequences of the crime to be important.

Annotation for NER was a complex task, as the process of the TRC hearings was flexible and witnesses could change without proper announcement. Each witness could also talk about multiple incidents involving different victims or a single incident with multiple victims. For the purpose of annotation, any person who is testifying in front of the Commission is considered a witness, not including eye-witnesses that are mentioned, but do not appear in front of the Commission. A victim is any person mentioned in the text who has been harmed, mentally or physically, as long as their case falls roughly into the jurisdiction of the TRC. Because this annotation is intended for NER, only people whose names are known can be included, even though other unnamed victims might be mentioned in the text.

A.2 Methods

A.2.1 Summarization Prompt

```
You are an academic assistant who summarizes historical witness reports. Your summaries should be concise, factual, and neutral in tone.

Write a short summary of the given witness report. Describe the course of events of the alleged crime and mention the names of the victims and witnesses.
```

A.2.2 NER Prompt

```
You are an academic assistant analyzing historical witness reports about human rights violations.

The given text is a witness statement about human rights violations during the South African apartheid.

Your task is to extract and classify the most important names of people that you find in the text.
Victims: The person that the witness is talking about and who was harmed
Witnesses: The person who is telling their story in front of the commission

Give your answers in JSON formatting with the fields "victims" and "witnesses". If the name does not appear in the text, write "NULL".
```

A.2.3 Violation Classification Prompt

```
You are an academic assistant analyzing historical witness reports about human rights violations.

You are given a witness report. Your task is to identify which type of crime the witness is talking about.

Choose a small number of classes from the following list. Only choose the classes when you are sure that they apply to the witness report. Only use crimes that are defined in the list and keep the exact same names.

Possible crime types:
* **killing**: someone was killed (intentionally or unintentionally)
* **torture**: pain inflicted
```

```
by state officials for coercion, information gathering or discrimination
* **detention**: victim was detained unlawfully or under unreasonably poor conditions
* **serious injuries**: victim has lasting health problems from the incident
* **assassination**: targeted, intentional killing for political reasons (usually by special forces soldiers)
* **shooting**: the victim was shot and hurt (may be deadly or not)
* **disappearance**: the victim has disappeared and it is currently unknown if they are dead or alive
* **context statement**: no specific incident, only general context
* **assault**: someone was assaulted and injured
* **harassment**: only use this if the harassment is severe (unlawful house search, death threats etc.) or if it is the main point of the hearing
* **bombing**: there was a bombing
* **destruction of property**: any form of vandalism that is not arson
* **arson**: someone set something on fire (but not human burning)
* **human burning**: a human was set on fire (deadly or not)
* **blackmail**
* **theft**

Output format: Comma-separated list of crime types.
```

A.2.4 Hyperparameter Tuning for Summarization

.

The only two models that were sensitive to their hyperparameters were Pegasus and TinyLlama. For Pegasus, increasing the number of beams for beam search and restricting the early stopping conditions resulted in summaries that were clearly more abstractive, contained fewer hallucinations, and were closer in length to the reference summaries.

For TinyLlama, the default parameters generated a single sentence with almost no relevance, which was repeated multiple times. Improved beam search parameters, combined with strong penalties on repetition, forced the model to produce coherent summaries. A low temperature setting reduced hallucinations at the cost of abstraction, resulting in improved numerical scores on the validation set. The exact parameters can be found in table 4.

A.2.5 Additional Experiments for Summarization

Limited Input: The process of successively summarizing paragraphs of texts seems intuitive to humans. Still, for a neural model, these intermediate steps could pose a problem, as contextual information between the chunks is lost. Since the model does not retain memory of the whole text when creating the final summary, errors in the intermediate summaries are easily propagated. It therefore seems as if models with a long context window have a clear advantage over models with a shorter context window, as they do not have to split the text. To test this hypothesis, we conducted a second experiment in which the maximum input size was reduced to 1024 tokens for all models. In this experiment, larger models like Mistral could still benefit from their superior context understanding abilities, but had to deal with the potential drawbacks of splitting and rejoining the texts like the other models.

Limited Output: Another potential source of error in the summarization process is the possibility that the joined intermediate summaries are too long to be processed, causing the model to repeat the process recursively. This means that the model processes the summary more than once, which can have a good or bad impact on the summary quality – in the best case, the model successively removes less important information, in the worst case, errors and hallucinations are propagated and amplified each time the text is passed to the model. Limiting the length of the intermediate summaries would reduce the risk of a document being processed multiple times. While limiting the output works well for the smaller, summarization-tuned models, the LLMs lack an encoder and predict the summary token by token, which means they cannot be properly limited to a specific length. We resorted to

Table 4: Hyperparameters: T5 and BART use default parameters from the Huggingface summarization pipeline, Pegasus and TinyLlama were fine-tuned, and all other models use default parameters.

Parameter	T5	BART	Pegasus	TinyLlama	Others
early_stopping	False	True	True	True	False
num_beams	1	4	5	4	4
temperature	1.0	1.0	0.6	0.3	0.8
top_p	1.0	1.0	1.0	0.8	0.9
top_k	50	50	50	50	40
repetition_penalty	1.0	1.0	1.0	1.2	1.1
length_penalty	1.0	2.0	2.0	2.0	1.0
no_repeat_ngram_size	0	3	0	3	0

truncating the generated intermediate summaries to 100 tokens, which might cause some critical information to be lost. This third experiment will determine whether the loss of information outweighs the potential benefits from reduced recursion.

A.2.6 Additional Experiments for NER

LLM Combination: For NER, the combination of the results for each chunk is not as straightforward as for summarization. The naive approach of concatenating the lists for victims and witnesses, respectively, resulted in too many duplicates. Removing duplicates is complicated since the context is lost – for example, "Paul" and "Mr. Smith" may or may not refer to the same person. For most experiments, I removed duplicates by fuzzy matching with token set ratio, a similarity measure that is robust to missing words, such as middle names or initials. This method works for all cases in which the full name of a person was detected at least once; however, it is also error-prone, as for example "Paul Smith" would match "Mrs. Smith". In a separate experiment, the LLM was tasked with combining the results independently. While this relies on the LLM's understanding of the JSON format and is prone to hallucinations, it can capture some details that fuzzy matching cannot.

Pipeline Approach: The length of the TRC documents is one of the main challenges for the LLM. Most of the information given in the text is irrelevant for the task, and only very few tokens determine whether or not a word should be included in the name list. Summarizing the input before applying NER could improve the results, as it simplifies the task for the LLM. We tested this hypothesis using human-written reference summaries to determine if summarization in general improves the results. We also performed NER with the result summaries from the first task to assess whether an LLM can complete the task fully on its own. An overview of all experiments can be seen in Table 5.

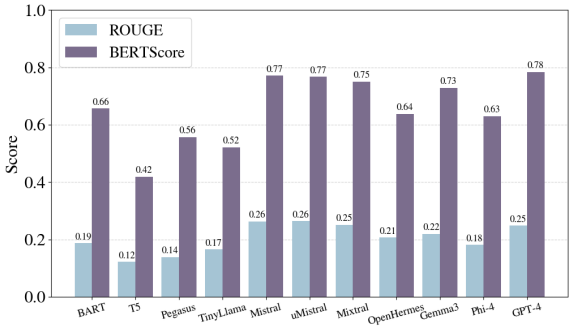

Figure 4: Summarization: Model Comparison with BERTScore and ROUGE-score.

A.3 Results

A.3.1 Additional Results for Summarization

Limited Input: When limiting the maximal input length to 1024, the setup for BART and Pegasus remained unchanged, as their context window is already only 1024 tokens long. For these models, the minimal changes between experiments are only due to the randomness of the model. For most of the other models, limiting the input did result in slightly lower summary quality, especially for Mistral and T5. For OpenHermes, however, limiting the input length improved the quality of the summary. Results like this can occur when a model's short context summarization abilities are significantly better than its long context summarization abilities, thereby overcoming the disadvantages introduced by chunk processing. Overall, the differences are small with a relatively high standard deviation. It can therefore be concluded that the advantages and disadvantages of chunk processing are balanced, confirming the effectiveness of this method in dealing with long inputs in summariza-

Table 5: Overview of different experiments.

Experiment	Task	Description
Main experiment	All	main results
Limited Input	All	input max. 1024 tokens, better model comparison
Limited Output	Sum.	intermediate output max. 100 tokens, reduced risk of recursion
LLM Combination	NER	JSON results combined by LLM
Reference Pipeline	NER	NER on reference summary
Pipeline Approach	NER	NER on LLM-generated summary

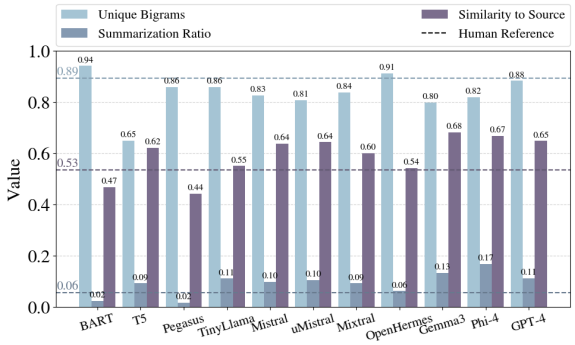

Figure 5: Summarization: Model Comparison with reference-free metrics.

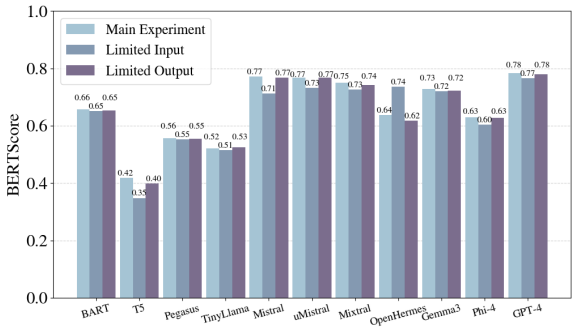

Figure 6: Summarization: BERTScore Results for the main experiment in comparison to the results with limited input and output.

tion.

Limited Output: When limiting the maximum output length of the intermediate summaries to 100 tokens, the models once again demonstrate surprising robustness. The performance decrease introduced by truncating intermediate summaries and potentially losing important information is minimal. The overall small changes in this experiment show that multiple passes over a long document are not necessarily a disadvantage, and that small truncations in the intermediate summaries do not influence the summary quality much.

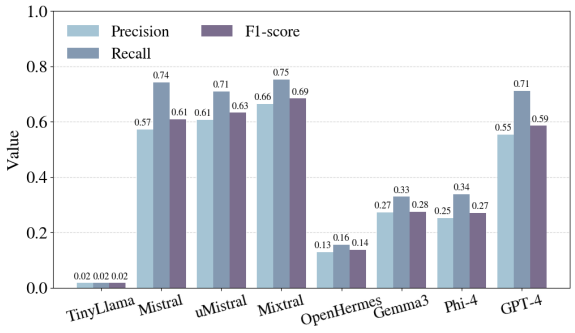

Figure 7: NER: Precision, recall and F1-score for each model, averaging over victims and witnesses.

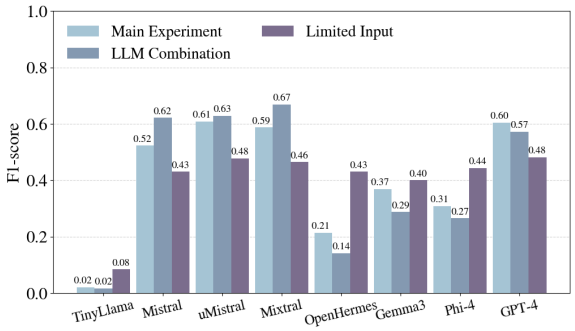

Figure 8: NER: Results for additional experiments with LLM Combination and Limited Input.

A.3.2 Additional Results for NER

Limited Input: For NER, the result quality is more dependent on the chunking process than for summarization. The Mistral models and Mixtral suffer from the limited input, while models that previously did not perform very well, like OpenHermes and Phi-4, improve. Even Gemma-3, with its long context, seems to improve when the information is presented in small chunks. For the better-performing models, the decrease in performance is probably due to the loss of context in between chunks. This problem is more prevalent for NER than for summarization, because for NER, there are very specific parts of the text that are relevant. If these parts are not contained in a chunk, the model

struggles with prediction and hallucinates. The robustness of summarization towards the chunking process is also one of the reasons for the success of the pipeline approach.

LLM Combination: For the better-performing models like Mistral and Mixtral, letting the model combine the intermediate results itself lead to improved results. Meanwhile, lower-performing models did not outperform the algorithmic approach of fuzzy matching.

A.4 Error Analysis

A.4.1 Summarization Reference

This is the reference summary referred to in Section 6.1.

> "Three witnesses testify about the death of Peter Mabilo, called Lathli or Luthli, who was shot and killed by police at his grandmother's house. The three witnesses are his mother, Mary Mabilo, his sister Florence Matlakala, and his grandmother Sheila Mabozo, who is also an eyewitness to the shooting. Peter Mabilo was a student and an active member of the ANC. He was arrested twice in 1985 and spent at first 25 days, then 8 months in prison. He was severely beaten by police before all charges were dropped, and he was released. After that, he fled the country without telling his family where he went. He returned in August 1987, when he lived with his grandmother and visited his mother, sister, and girlfriend for a day. The next day, police came to Mrs. Mabozo's house and ordered the family to come outside. Mr. Mabilo was hiding inside the house, but surrendered after the police threw teargas inside. He was then shot in the hallway of the house. Mrs. Mabozo recalls that the furniture was destroyed and that there were bullet holes in the walls. The house of Mr. Mabilo's mother and sister was searched afterwards, and the family was harassed by the police. Mr. Mabilo's funeral was interrupted by police, who only gave them 5 minutes to bury Mr. Mabilo. Mrs. Mabilo accuses a policeman named Terra Cacamba of killing her son."

A.4.2 Detailed Error Analysis for Summarization Example

In the following quotations of the model's summaries of this case, hallucinations are highlighted in red, highly irrelevant facts are highlighted in blue, and important facts are highlighted in green.

> 1.:ADV POTGIETER: Okay, then we'll stand down - stand your evidence down for the moment and I think we'll - will take the evidence of your daughter, of Florence and we could come back to you if there's something you want to add to what you've said..ADV POTGIETER: Okay, then we'll stand down - stand your evidence down for the moment and I think we'll - will take the evidence of your daughter, of Florence and we could come back to you if there's something you want to add to what you've said.
> – T5, BERTScore 0.1896, ROUGE 0.0777

This summary is highly extractive, as were most summaries that T5 created. This is likely due to the training on news articles, where not much rephrasing is necessary for summarization. In some cases, the extractive summaries contained important information, although in this example, it merely quotes a moment from the middle of the testimony where procedural matters are discussed.

> Mary Mabilo is the mother of Lathli Mabilo .<n>Lathli was shot dead in 1987 at the home of Ms Sheila Mabozo .<n>Ms Mabilo says five of the accused were arrested in 1985 and spent 25 days in jail .<n>Misty Mabozo was found dead at her home in Ipopeng .<n>She was allegedly murdered by Terra, who is on trial for her murder .<n>Mary Mabilo told the court how he was told Terra had killed his son .
> – Pegasus, BERTScore 0.6502, ROUGE 0.1825

Pegasus' summaries are much more abstractive. Each sentence can still be traced back to a sentence in the original text, but the model was able to abstract from the dialogue format of the original text by adding names and adapting the pronouns. The information in the text is mostly correct, but not always important and coherent. For example, the sentence about the five accused who were arrested

refers to the victim's first arrest before his exile, which is not clear from the summary.

In addition, Pegasus' summaries are usually very short, and while they include important details in the summary, they are often incomprehensible due to missing context or too little paraphrasing. In this case, the model also switched the victim's name to "Misty Mabilo", a name that appears in the original text only once in reference to one of the witnesses. This is an example of a minor error that can significantly misrepresent the witness report, rendering the summary inaccurate and useless. Most models included small but significant errors like this in their summaries.

> Mary Mabilo, Florence Matlakala and Sheila Mabozo give evidence. All three are from Galishewe Village in Mpumalanga province. Mary tells how her son, Lathli, was shot dead in 1987. Florence tells how five of her children were arrested and held for eight months. Sheila says she was blindfolded and had sharp instruments put into her ears.
> – BART, BERTScore 0.6508, ROUGE 0.1202

While BART sometimes has issues with over-extractiveness, too, in this example, it abstracts reasonably well. The information is mostly correct, but important details of the main incident are missing. In contrast to the reference summary, this summary follows the temporal events of the hearing (one witness after the other), instead of combining the testimonies into one description of the course of events. The inability to understand the order of events was a common factor among all models. In the case of the non-LLM models, this can be attributed to the chunking process, which makes it especially challenging to maintain context over long texts and hinders the model's ability to, for example, recognize that witnesses are describing the same incident. While this summary is objectively better than the Pegasus summary, it has the same BERTScore and a lower ROUGE score due to more abstraction.

> The Commission has concluded that there is insufficient evidence to charge the accused with the crime of murder. The evidence presented by the prosecution does not meet the standard of proof required for a conviction of murder, and there is no evidence of a motive or premeditation in relation to the crime. The investigation into the incident is ongoing, and the Commission will continue to monitor the progress of the investigation to determine whether any further charges should be brought against the accused. This entry was posted in Uncategorized and tagged murder, South Africa. Bookmark the permalink. Post a comment or leave a trackback: Trackback URL.
> – TinyLlama, BERTScore 0.4590, ROUGE 0.1342

TinyLlama occasionally provides relatively good summaries. They mostly suffered from the omission of important details, which were replaced by well-formulated, but generic and meaningless sentences. This summary is an extreme example, where the model's entire answer was comprised of hallucinated or meaningless sentences. No trace of the original text remains, except for the fact that there was a murder. This problem of over-abstraction into meaninglessness is the exact opposite of the issues that the encoder-decoder models struggled with, and a result of the architecture. Without a decoder, the LLM cannot read the complete text and then write a full summary; instead, it simply predicts the next most likely token. Without a very high number of parameters and sufficient training data and instruction fine-tuning, this results in many generic sentences.

> The witnesses, Mary Mabilo, Florence Matlakala, and Sheila Mabozo, testified about the shooting of Mary Mabilo's son, Lathli Mabilo, in 1987. Lathli had been missing for more than a year before his death. He was wanted by the police and was frequently harassed. He was killed at Sheila Mabozo's house in Ipopeng, South Africa. The police interrogated Mary Mabilo about the incident, and two women came to her house and took her furniture without her consent. Ms Matlakala testified that her brother, Lathli, returned home in 1987 but was killed on the same day. The police harassed the family before and during the funeral. Ms Mabozo, Lathli's grandmother, was interrogated by the police, blindfolded, and her ears were damaged with sharp objects. She was held in the Barkley cells

for two days. Luthli, who had been living with Ms Mabozo for a month, was shot and killed by the police outside their house. The police had previously shot tear gas into the house. Ms Mabilo, Luthli's mother, also testified and asked for Terra, a policeman, to be called to explain why he killed her son. Ms Mabilo is a poor woman who
– uMistral, BERTScore 0.7854, ROUGE 0.3012

This is a subjectively good summary of the incident. The information is correct. Small details that could be considered significant, such as the victim's previous arrests, are omitted. The order of events is accurate, but described in a somewhat confusing manner, and the structure of the original document (one witness after the other) is still recognizable.

[...] Ms Mabozo, Luthli's grandmother, provided additional details about the shooting itself. She stated that after the police took Luthli's father away, they were told to take the children outside. Luthli was inside the house when the police shot him. The family found the house damaged and full of bullet holes when they returned. [...]
In conclusion, the Commission expressed its sympathy for the family and vowed to investigate the circumstances surrounding Luthli's death [...].
– Mistral, BERTScore 0.7143, ROUGE 0.2470

In this example, the difference between Mistral and uMistral is evident to a human reader, which was not always the case. In contrast to Mistral, uMistral produced more hallucinations. The last, highly irrelevant paragraph can be seen as an example of a generic statement that was generated because the system prompt encourages the model to be polite and eloquent. Even though both models were run with the same hyperparameters, Mistral seems to abstract more.

The witness report contains statements from three witnesses, Mary Mabilo, Florence Matlakala, and Sheila Mabozo, regarding the shooting and death of Mary's son, Lathli Mabilo, in Kimberley, South Africa in 1987. Mary, Lathli's mother, states that Lathli was involved in anti-apartheid political activities and had been arrested multiple times. She describes how Lathli had been missing for a year and seven months before she saw him again for one day, and then he was killed. Mary suspected that "those who were after him" were responsible for his death. Florence, Lathli's sister, testified that Lathli had returned to Kimberley to deliver a message to someone and that she took him to meet his girlfriend before he planned to leave town again. She describes how the family was harassed by the police before and after Lathli's funeral, which was attended by his organization and family. Sheila, Lathli's grandmother, testifies that Lathli was shot outside her home while he had his hands up, and that the police shot him multiple times and then entered the house and continued to shoot him. The witnesses request that the Commission investigate the circumstances surrounding Lathli's death and the identity of the police officer "Terra" who they believe was involved in his killing.
– Mixtral, BERTScore 0.7092, ROUGE 0.2653

Even though this summary has a lower BERTScore than both Mistral models, it is objectively the best summary of the case. Although it remains structurally similar to the original document, the temporal order of events is now comprehensible. It is a long summary, and some of the details may not be crucial, but the information provided is accurate and covers nearly all the important facts. The lower scoring of this summary means that it is more different from the reference summary without necessarily being worse. This indicates that the evaluation metrics, although providing a good overview when averaging over multiple texts, cannot fully assess the quality of a single summary. Very interestingly, Mixtral tends to capture details that other models miss, in this case, for example, the legally important fact that the victim had his hands up when he was shot.