Non-Archival Abstract

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
University of Konstanz
miriam.schirmer@uni-konstanz.de

In South Africa, the Truth and Reconciliation Commission (TRC) [1] produced an extensive collection of testimonies documenting the experiences of victims and witnesses of apartheid. While this data could serve as a valuable historical and legal resource, it is difficult to access for a wider audience due to its volume, unstructured format, and linguistic complexity. We examine whether open-source LLMs can extract key information from such complex texts. In a comparative analysis, we focus on summarization and NER of victim and witness names, as well as the identification of the violations discussed. This combination of tasks is particularly interesting, as recent studies show strong performance of LLMs on abstractive summarization (Liu and Healey, 2026), while their effectiveness for named entity recognition remains less consistent (Wang et al., 2025). We also introduce a pipeline approach that improves the results by summarizing the reports before performing NER.

Methods The dataset used for this study contained 200 transcripts describing human rights violations during apartheid. To obtain a human baseline, 3 experts annotated the dataset, following detailed annotation guidelines. We evaluated 7 models for summarization and fine-grained NER, more specifically the extraction of personal names of witnesses and victims, as well as a label for the crime that was committed (e.g., "shooting", "torture"). The open-source models were TinyLlama (Touvron et al., 2023), Mistral (Jiang et al., 2023), Mixtral (Jiang et al., 2024), OpenHermes (Teknium, 2023), Gemma3 (Mesnard et al., 2024), and Phi-4 (Abdin et al., 2024). A second version of Mistral, which we call uMistral, received unstructured prompts to disable parts of the instruction tuning. In addition, we included GPT-4o-mini (OpenAI, 2024) as a comparison to commercial models. As baseline models for summarization, we also added T5 (Raffel et al., 2019), BART (Lewis et al., 2019), and Pegasus (Zhang et al., 2019). For victim and witness detection, the baseline was SpaCy (Honnibal et al., 2020). For all prompted models, the prompt template was identical and systematically engineered. Due to the limited context length of the models, reports sometimes had to be split into chunks on a sentence level. The results were concatenated and passed to the model again to create a single output. In addition, we improved the results with a pipeline approach, where we applied NER on the model-generated summaries to assess whether summarization simplifies NER for long TRC documents. The results are evaluated using BERTScore for summarization and micro F1-score for NER.

Results The results ranged from moderate to strong in both tasks (see Table 1). Most models clearly outperformed the summarization baseline, with Mistral reaching the highest BERTscore of 0.77. Performance on the NER task was comparatively lower, with uMistral achieving the best result (F1-score 0.61). The performance of the SpaCy baseline was poor (F1-score 0.19), since non-LLM NER models cannot distinguish between roles without further fine-tuning. This shows that, despite the LLMs' ability to solve a variety of 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. The pipeline approach improved the results, especially for the models that previously showed lower performance. The strongest increase could be seen for OpenHermes, which improved by 0.23 in F1-score. Summarizing not only shortened the input, but also reformulated it to focus on the main events of the crime and relevant names, which helped the NER task.

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

Table 1: Evaluation metrics for summarization (BERTScore) and NER (micro F1)

Model	Sum. BERTScore	NER F1	Pipeline F1
SpaCy	-	0.1901	-
BART	0.6567	-	-
Pegasus	0.5574	-	-
T5	0.4186	-	-
TinyLlama	0.5218	0.0248	0.0258
Mistral	0.7708	0.5086	0.5685
uMistral	0.7673	0.6064	0.6058
Mixtral	0.7504	0.5814	0.5810
OpenHermes	0.6374	0.2569	0.4881
Gemma3	0.7292	0.3864	0.5975
Phi4	0.6304	0.3644	0.4675
GPT-4o-mini	0.7832	0.5916	0.6842

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.

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

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. 2024. Mixtral of experts. *Preprint*, arXiv:2401.04088.

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.

Shengjie Liu and Christopher G. Healey. 2026. Abstractive summarization of large document collections using gpt. *IEEE Transactions on Big Data*, pages 1–14.

Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, and 88 others. 2024. Gemma: Open models based on gemini research and technology. *Preprint*, arXiv:2403.08295.

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

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.

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.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, Guoyin Wang, and Chen Guo. 2025. GPT-NER: Named entity recognition via large language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4257–4275, Albuquerque, New Mexico. Association for Computational Linguistics.

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.