Knowledge Distillation in Compact Models: An Approach Applied to Text Processing for Public Security

Ricardo Rodrigues Barcelar[1], **Leonardo Arruda Vilela Garcia**[1], **Alan Papafanurakis Heleno**[1], **Thiago Meirelles Ventura**[1], **Allan Gonçalves de Oliveira**[1]

[1]Instituto de Computação – Universidade Federal de Mato Grosso (UFMT)
78060-900 – Cuiabá – MT – Brasil

{ricardo.barcelar, leonardo.garcia1, alan.heleno}@sou.ufmt.br,
{thiago, allan}@ic.ufmt.br

Abstract. *This article aims to develop a model for summarizing police reports in the context of public security, with a focus on execution on limited hardware. The approach combined hybrid distillation (logits and intermediate representations) and supervised fine-tuning with LoRA, using a corpus of 19,286 police reports. The evaluation was conducted using automatic metrics (BERTScore-F1) and qualitative analysis by specialists. The results demonstrate that the distilled model generates clear, coherent, and semantically appropriate summaries, comparable to those of the larger model, with superior computational performance, including in hardware-constrained environments.*

1. Introduction

Large Language Models (LLMs) have significantly transformed the field of Natural Language Processing (NLP), expanding the capabilities for processing complex texts. With a transformer-based architecture and the number of model parameters increasing from hundreds of millions to hundreds of billions in recent years [Zheng et al. 2024], this has allowed for advancements in tasks such as summarization, classification, and information extraction [Brown et al. 2020].

However, the increase in model size raises the cost of inference, restricting their application in operational scenarios with limited hardware. Knowledge distillation, a compression technique in which a smaller model (student) is trained to reproduce the behavior of a larger model (teacher) by aligning outputs (logits), intermediate representations (features), or both, emerges as an efficient alternative for developing compact models [Hinton et al. 2014, Jiao et al. 2020].

Nevertheless, excessive discrepancies between the capabilities of the teacher and student models can create informational bottlenecks, compromising performance. Thus, optimized specialization workflows are necessary to balance capacity, performance, and computational efficiency [Setiawan 2024, Zhang et al. 2023].

Although training demands high computational cost, their use as base models has enabled adaptations to specific contexts, such as public security [Sarzaeim et al. 2024]. In this context, police reports (in Brazil, called ”boletim de ocorrência”), documents that textually record potentially criminal events, present varied terminology, complex contextualizations, and a lack of standardization, hindering automated processing. The manual analysis of these texts consumes time and human resources and is susceptible to inconsistencies, which positions LLMs as promising tools, including by demonstrating abilities such as in-context learning and instruction following [Wei et al. 2022].

While there are initiatives for the automated analysis of legal texts [Pereira et al. 2025, Ramesh et al. 2024], there is a significant gap in the application of LLMs in public security, especially for the automatic summarization of narratives, due to the complexity of the domain and the scarcity of annotated data. Therefore, this article proposes a knowledge distillation approach for summarizing police reports, with the objective of developing a model capable of generating coherent, semantically accurate summaries aligned with the technical vocabulary of public security. The methodology integrates hybrid distillation (logits and intermediate representations) and supervised fine-tuning, aiming to maintain the robustness of the results even in environments with limited hardware, ensuring accuracy and conciseness in the generation of summaries.

The presentation of this work is organized as follows: section 2 presents related works, contextualizing existing research, providing a theoretical basis, and identifying gaps; section 3 describes the proposed methodology, including the architecture and the data used; section 4 presents the results and discussions; finally, section 5 presents the conclusions and future directions.

2. Related Works

The use of LLMs in public security has been gaining attention due to their ability to process large volumes of textual data and assist in analytical and operational tasks. Sarzaeim et al. [2024] investigated the application of LLMs in classification tasks for criminal data analysis, demonstrating their effectiveness in identifying emerging patterns that aid in the prevention of and response to illicit activities. However, the study is limited to supervised tasks.

The generation of textual content in institutions with complex processes was explored by Pereira et al. [2025] with the INACIA system, which automates analysis steps at the Federal Court of Accounts, including the production of textual documents. This work suggests the potential of LLMs for similar tasks in public security. However, the absence of specific studies on text generation in police contexts, especially with compact models, reveals a significant gap. Few works address the analysis of police documents, such as police reports.

Regarding the use of mechanisms for the preparation and creation of LLM models, knowledge distillation is a consolidated technique for creating compact and efficient models, reducing computational costs while maintaining competitive performance. Hinton et al. [2014] established the foundations of distillation, proposing the transfer of logits from a teacher model to a student, using a combined loss function that balances cross-entropy and Kullback-Leibler divergence. Jiao et al. [2020] leveraged this approach with TinyBERT, incorporating intermediate representations to improve the transfer of contextual knowledge, especially in natural language understanding tasks.

Several studies have reinforced the effectiveness of knowledge distillation in building compact and efficient models for Natural Language Processing tasks. Vakili et al. [2024] demonstrated that distilled versions of BERT maintain robust performance even on resource-constrained devices, aligning with the objective of the present study. Zhang et al. [2023] introduced IBKD, a method that employs the information bottleneck principle to distill textual representations without relying on logits. Park et al. [2021] proposed an approach based on contextual relationships between representations, achieving competitive results in language understanding tasks. Aguilar et al. [2020] highlighted that transferring intermediate layers can eliminate the need for softmax outputs, while Sun et al. [2019] emphasized the alignment of hidden states (intermediate representations) as a strategy to preserve the semantics and context modeling of the original model.

These contributions provide theoretical and practical foundations for the application of distillation in specific domains, requiring methodological adaptations to ensure efficiency and semantic fidelity. The specialization of LLMs for specific domains often involves fine-tuning, whether supervised or unsupervised. Ramesh et al. [2024] explored efficient fine-tuning techniques, such as Parameter-Efficient Fine-Tuning (PEFT) and Quantized Low-Rank Adaptation (QLoRA), demonstrating precise and contextually appropriate responses in adjusted models. However, these approaches require labeled data, which can be a limitation in scenarios like public security, where labels are scarce.

Although the reviewed works demonstrate advances in the application of LLMs, the combination of knowledge distillation and fine-tuning for text generation tasks in public security remains underexplored. Most studies focus on classification or generic domains, neglecting the analysis of texts for public security forces and operation on modest hardware. This study fills this gap by seeking to enable practical applications in public security with computational and legal constraints.

3. Materials and Methods

This section describes the methodology used for the knowledge distillation of a large language model (Qwen 2.5-14B), previously trained for summarizing police reports, into a more compact model (Qwen 2.5-1.5B) by adopting a hybrid methodology combined with complementary fine-tuning. In this work, the Qwen 2.5-14B model will be referred to as the teacher and the Qwen 2.5-1.5B model as the student.

3.1. Dataset

The success of specialization or knowledge distillation strategies for an LLM depends on a structured set of text or speech (corpus) representative of the target domain. In the police context, police reports, which contain narratives of criminal events, fit the desired characteristics. In this regard, the data used in this research consists of 19,286 police reports written in Portuguese, previously anonymized, provided by the Polícia Judiciária Civil do Estado de Mato Grosso (PJC/MT), and made available under an institutional agreement for scientific research, in compliance with the General Data Protection Law (LGPD, Law No. 13,709/2018).

The Police Report is a document widely used by public security forces to record criminal events, containing data on those involved, locations, dates, criminal classification, etc., and unstructured information, such as the narratives - the attribute of interest in this research. These reports were extracted from the Internal System for Recording Police Incidents (SROP) used by PJC/MT and include records made in the year 2024 in the city of Cuiabá/MT, covering various types of incidents, ensuring the randomness of events and diversity of narratives.

3.2. Experiment Setup

All scripts were implemented in the Python programming language, version 3.10.12, using the Hugging Face Transformers, Unsloth, and Torch libraries as the main tools in a Jupyter Notebook environment for knowledge distillation and fine-tuning tasks. The inference and training of the models were carried out in a virtualized environment with 8 Intel Xeon Gold 6426Y processing cores (vCPU), 64GB of RAM, and an NVIDIA L40S 48GB VRAM GPU.

The experiments were conducted in 4 sequential stages, visually detailed in Figure 1, which included (i) the processing of the police reports, (ii) hybrid distillation combining the transfer of logits and intermediate representations, (iii) supervised fine-tuning specific to the summarization task, and (iv) the evaluation of the student model, which are described in the following sections.

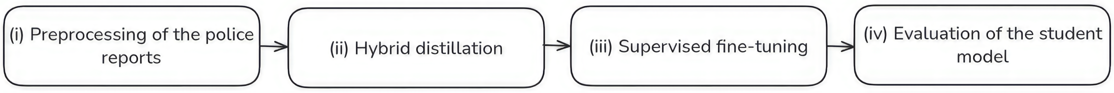
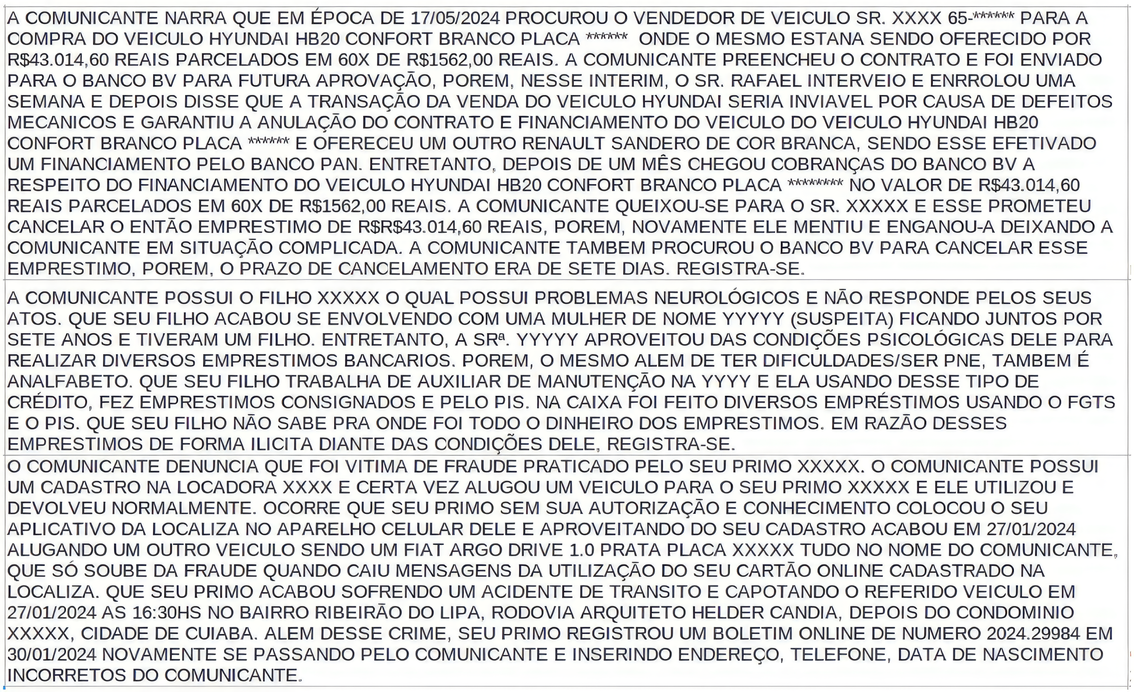

(i) Preprocessing of the police reports
(ii) Hybrid distillation
(iii) Supervised fine-tuning
(iv) Evaluation of the student model

Figure 1. Stages of experiment execution

Considering that the primary purpose of the models in this research is the ability to accurately and qualitatively summarize police reports, the narrative, as illustrated in Figure 2, was the main element in the construction of the training dataset. Police reports with very short (less than 50 words) or poorly formed narratives were kept so that the model could learn to handle this type of information.



A COMUNICANTE NARRA QUE EM ÉPOCA DE 17/05/2024 PROCUROU O VENDEDOR DE VEICULO SR. XXXX 65-****** PARA A COMPRA DO VEICULO HYUNDAI HB20 CONFORT BRANCO PLACA ****** ONDE O MESMO ESTANA SENDO OFERECIDO POR R$43.014,60 REAIS PARCELADOS EM 60X DE R$1562,00 REAIS. A COMUNICANTE PREENCHEU O CONTRATO E FOI ENVIADO PARA O BANCO BV PARA FUTURA APROVAÇÃO, POREM, NESSE INTERIM, O SR. RAFAEL INTERVEIO E ENRROLOU UMA SEMANA E DEPOIS DISSE QUE A TRANSAÇÃO DA VENDA DO VEICULO HYUNDAI SERIA INVIAVEL POR CAUSA DE DEFEITOS MECANICOS E GARANTIU A ANULAÇÃO DO CONTRATO E FINANCIAMENTO DO VEICULO DO VEICULO HYUNDAI HB20 CONFORT BRANCO PLACA ****** E OFERECEU UM OUTRO RENAULT SANDERO DE COR BRANCA, SENDO ESSE EFETIVADO UM FINANCIAMENTO PELO BANCO PAN. ENTRETANTO, DEPOIS DE UM MÊS CHEGOU COBRANÇAS DO BANCO BV A RESPEITO DO FINANCIAMENTO DO VEICULO HYUNDAI HB20 CONFORT BRANCO PLACA ******** NO VALOR DE R$43.014,60 REAIS PARCELADOS EM 60X DE R$1562,00 REAIS. A COMUNICANTE QUEIXOU-SE PARA O SR. XXXXX E ESSE PROMETEU CANCELAR O ENTÃO EMPRESTIMO DE RR43.014,60 REAIS, POREM, NOVAMENTE ELE MENTIU E ENGANOU-A DEIXANDO A COMUNICANTE EM SITUAÇÃO COMPLICADA. A COMUNICANTE TAMBEM PROCUROU O BANCO BV PARA CANCELAR ESSE EMPRESTIMO, POREM, O PRAZO DE CANCELAMENTO ERA DE SETE DIAS. REGISTRA-SE.
A COMUNICANTE POSSUI O FILHO XXXXX O QUAL POSSUI PROBLEMAS NEUROLÓGICOS E NÃO RESPONDE PELOS SEUS ATOS. QUE SEU FILHO ACABOU SE ENVOLVENDO COM UMA MULHER DE NOME YYYYY (SUSPEITA) FICANDO JUNTOS POR SETE ANOS E TIVERAM UM FILHO. ENTRETANTO, A SRª. YYYYY APROVEITOU DAS CONDIÇÕES PSICOLÓGICAS DELE PARA REALIZAR DIVERSOS EMPRESTIMOS BANCARIOS. POREM, O MESMO ALEM DE TER DIFICULDADES/SER PNE, TAMBEM É ANALFABETO. QUE SEU FILHO TRABALHA DE AUXILIAR DE MANUTENÇÃO NA YYYY E ELA USANDO DESSE TIPO DE CRÉDITO, FEZ EMPRESTIMOS CONSIGNADOS E PELO PIS. NA CAIXA FOI FEITO DIVERSOS EMPRÉSTIMOS USANDO O FGTS E O PIS. QUE SEU FILHO NÃO SABE PRA ONDE FOI TODO O DINHEIRO DOS EMPRESTIMOS. EM RAZÃO DESSES EMPRESTIMOS DE FORMA ILICITA DIANTE DAS CONDIÇÕES DELE, REGISTRA-SE.
O COMUNICANTE DENUNCIA QUE FOI VITIMA DE FRAUDE PRATICADO PELO SEU PRIMO XXXXX. O COMUNICANTE POSSUI UM CADASTRO NA LOCADORA XXXX E CERTA VEZ ALUGOU UM VEICULO PARA O SEU PRIMO XXXXX E ELE UTILIZOU E DEVOLVEU NORMALMENTE. OCORRE QUE SEU PRIMO SEM SUA AUTORIZAÇÃO E CONHECIMENTO COLOCOU O SEU APLICATIVO DA LOCALIZA NO APARELHO CELULAR DELE E APROVEITANDO DO SEU CADASTRO ACABOU EM 27/01/2024 ALUGANDO UM OUTRO VEICULO SENDO UM FIAT ARGO DRIVE 1.0 PRATA PLACA XXXXX TUDO NO NOME DO COMUNICANTE, QUE SÓ SOUBE DA FRAUDE QUANDO CAIU MENSAGENS DA UTILIZAÇÃO DO SEU CARTÃO ONLINE CADASTRADO NA LOCALIZA. QUE SEU PRIMO ACABOU SOFRENDO UM ACIDENTE DE TRANSITO E CAPOTANDO O REFERIDO VEICULO EM 27/01/2024 AS 16:30HS NO BAIRRO RIBEIRÃO DO LIPA, RODOVIA ARQUITETO HELDER CANDIA, DEPOIS DO CONDOMINIO XXXXX, CIDADE DE CUIABA. ALEM DESSE CRIME, SEU PRIMO REGISTROU UM BOLETIM ONLINE DE NUMERO 2024.29984 EM 30/01/2024 NOVAMENTE SE PASSANDO PELO COMUNICANTE E INSERINDO ENDEREÇO, TELEFONE, DATA DE NASCIMENTO INCORRETOS DO COMUNICANTE.

Figure 2. Examples of police report narratives

For the knowledge distillation task, a stratified set of 8,502 previously anonymized police reports was used to build the dataset. This dataset was enriched with an additional column containing summaries produced by the teacher model and subsequently partitioned into 80% for training and 20% for validation, allowing for the evaluation of the model during the knowledge distillation process. Additionally, for the supervised fine-tuning task, a second dataset was created with 10,284 police reports, following the same enrichment and division process into training and validation sets. To perform the final evaluation of the models, the remaining 500 police reports were set aside to compose the test dataset.

The Qwen 2.5-14B model (teacher) was selected owing to Portuguese pre-training and the availability of an architecturally compatible compact variant (Qwen 2.5-1.5B), which supported the viability of hybrid distillation (logits and intermediate features) within the Transformers/Unsloth pipeline. Prior to the experiments, the teacher had been trained to summarize police report narratives. Hybrid knowledge distillation of logits and intermediate features was applied to the Qwen 2.5-1.5B model (student) using the Unsloth library.

Considering that the objective of distillation is to reproduce the teacher's behavior at the output, or to capture its way of processing information internally, it was verified through several rounds of tests and parameter modifications that logits distillation alone would not be sufficient for summarization tasks, leading to the choice of hybrid distillation of logits and features. This forces the student model to internalize the teacher's intermediate representations (features), which proved to be significant for the intended task, where the model's structure and internal processing influence the outcome. In this context, to control the weight of each part of the learning, the parameters alpha=1.0, beta=4.0, use_logits=True, use_features=True, and temperature=0.8 were used, where alpha represents the weight assigned to logits distillation and beta to features.

The training parameter configuration aligns with the proposal of Sun et al. [2019], in which the use of weighted losses on the teacher's hidden states (intermediate representations) is fundamental for complex tasks, especially when there is a large discrepancy between the sizes of the teacher and student models. This training was executed over 3 epochs, with the model in Mixed Precision mode (bfloat16), Adamw_torch optimizer on a dataset prepared as described in the previous section.

Although distillation transfers knowledge from the teacher model to the student model, this process does not, by itself, guarantee the student's maximum adaptation to the target task [Jiao et al. 2020]. Therefore, after the distillation step, additional supervised fine-tuning was performed using the LoRA (Low-Rank Adaptation of Weights) method, with the objective of adjusting the student model to the particularities of the summarization task. With LoRA, only a subset of the model's linear projections was updated, keeping the other weights frozen. This training was executed over 3 epochs, with the model in Mixed Precision mode (bfloat16), Adamw_torch optimizer, and a learning rate of 2e-4.

3.3. Evaluation Metrics

The models were tested using a set of 500 police reports. For each police report, a diversified prompt was constructed, instructing the model to perform a summary similar to the following examples:

- Summarize the incident concisely, with an emphasis on the dynamics of the event and those involved.
- Produce a descriptive summary of up to 100 words, highlighting the dynamics of the facts and those involved.
- Synthesize the narrative in up to 50 words, with a focus on clarity and central elements.

The evaluation of the results was carried out using BERTScore-F1, an automatic metric that measures the semantic similarity between the generated text and a reference, based on contextual embeddings from BERT [Zhang et al. 2020]. This metric is especially suitable for identifying textual variations that preserve the same meaning, appropriate for evaluating the semantic precision of the produced summaries. BERTScore-F1 was used to compare the performances of the teacher model and the student model, allowing for an automated analysis of the quality of the generated summaries.

Additionally, a qualitative evaluation was applied in which five civil police officers (experts) were invited to evaluate, according to a Likert scale (1 to 5), samples of the texts generated by the distilled model, assessing:

- Clarity: Readability and structure of the text;
- Narrative coherence: Logical consistency of the text and;
- Operational utility: Relevance for police decisions.

From the police reports in the test set, 40% were set aside to be evaluated by experts, who assessed each summary with the corresponding original report as reference. Each expert evaluated 40 summaries, 20 generated by the teacher model and 20 by the student model. This design aimed not only to evaluate the summaries produced by the student but also to measure the quality of the teacher's generation, enabling a comparative quality metric for the summaries of both models.

4. Results and Discussion

The application of the proposed methodology resulted in a series of findings related to the quality of the generated summaries, the semantic similarity with the reference model, and the computational performance in scenarios with limited resources. This section aims to present these results in a structured manner, covering automatic and qualitative evaluations, and operational metrics, as well as discussions on their implications for the use of compact LLMs in the context of public security.

4.1. Automatic Evaluation

The quantitative analysis, conducted based on the BERTScore-F1 metric, demonstrated high semantic similarity between the summaries produced by the teacher (Qwen 2.5-14B) and student (Qwen 2.5-1.5B) models. The technique evaluated the similarity between texts based on the harmonic mean of precision and recall of cosine similarity between token embeddings, allowing for the capture of semantic nuances, even in the presence of lexical variations. The results obtained from the analysis of 500 samples are described in Table 1.

The results indicate that, on average, the student model was able to significantly reproduce the semantic content of the teacher model. The high median value (0.85) and

75th percentile (0.87) reinforces the consistency of performance, despite the variability observed in certain cases (minimum of 0.78).

Table 1. BERTScore-F1 results for summary comparison

Statistical Metric	Value (0-1)
Mean	0.85
Standard Deviation	0.03
Minimum	0.78
25th Percentile	0.84
Median (50%)	0.85
75th Percentile	0.87
Maximum	0.91

4.2. Expert-Based Qualitative Evaluation

In the qualitative evaluation, specialist civil police officers directly compared the summaries generated by the teacher model and the student model based on criteria of clarity of the summarized text, coherence with the original narrative, and operational utility for police officers. The results are summarized in Table 2.

Table 2. Qualitative evaluation of summaries (Likert Scale 1-5)

Criterion	Teacher Average (1-5)	Student Average (1-5)
Clarity	4.75	4.35
Coherence	4.61	4.20
Operational Usefulness	4.09	3.92

Regarding clarity, the summaries produced by the student model received an average score of 4.35, slightly below the 4.75 assigned to the teacher model. This result suggests that the structure and readability of the text were well assimilated by the student model, indicating that formal and linguistic aspects were satisfactorily transferred during the distillation process.

In terms of narrative coherence, the difference observed was also modest (4.61 for the teacher versus 4.20 for the student). The student model was perceived as capable of maintaining an acceptable narrative coherence, reproducing the internal logic learned from the teacher, albeit with minor errors or simplifications. This indicates that the ability to construct coherent narratives was inherited, with room for improvement.

With respect to operational usefulness, the results suggest that the student model did not always manage to extract and condense the most relevant elements for practical police use. Nevertheless, the score still falls within a positive range (”Good” on the Likert scale), indicating that knowledge transfer did occur, although with some loss of critical information.

4.3. Computational Performance Evaluation

The computational performance evaluation was conducted using 50 police reports from the evaluating set, with which a prompt for summarization was assembled. The teacher model showed an average inference time of 6.26 seconds per document, with a throughput of 69.73 tokens/s and a maximum VRAM usage of approximately 28.6 GB. In contrast, the student model processed the documents in an average time of 3.42 seconds, with a throughput of 125.59 tokens/s, consuming only 3.07 GB of VRAM. Both models processed the same samples, demonstrating consistency in the outputs.

4.4. Discussion

It was observed in the initial tests of the experiments that, due to the significant difference in capacity between the models (in the case of a student with a very reduced number of parameters), there was an internalization of general representations and patterns from the teacher without, however, fully optimizing its ability to summarize texts. This confirmed the need for additional fine-tuning, aiming to specialize the student model in the target task by adjusting its parameters based on supervised examples and, thus, improving the generation of summaries.

The results obtained demonstrate the viability of hybrid distillation combined with supervised fine-tuning for producing summaries of police reports by smaller models. The distilled model maintained high semantic adherence to the teacher model's responses, as proven by the BERTScore-F1, which is relevant in real-world public safety applications that require precision in capturing the reported facts.

The proximity of the qualitative results indicates that distillation preserved essential characteristics of the teacher, such as clarity and coherence, despite the reduction in parameters. On a 1–5 Likert scale, the student scored 4.35 vs. 4.75 in clarity (-0.40; $\approx -8.4\%$), 4.20 vs. 4.61 in coherence (-0.41; $\approx -8.9\%$), and 3.92 vs. 4.09 in operational usefulness (-0.17; $\approx -4.2\%$) (Table 2). These differences suggest a modest loss in complex summaries but do not materially compromise the final performance for the intended use, given the high semantic alignment between models and the efficiency gains already reported. Importantly, operational usefulness received the lowest absolute ratings for both models, indicating weaker performance on this criterion overall; this likely reflects the reliance on generic prompts that did not encode role-specific operational requirements.

The computational analysis revealed substantial efficiency gains provided by distillation. The model, distilled and adjusted via fine-tuning, was able to generate summaries nearly twice as fast and with an approximately 90% reduction in GPU memory consumption when compared to the original 14-billion-parameter model. This expressive difference demonstrates that the student model is significantly lighter and more agile, making it more viable and agile for practical applications in computationally constrained environments.

All code related to this work is publicly available at `https://doi.org/10.5281/zenodo.16877253`

5. Final Considerations

This work presented a knowledge distillation approach for summarizing police reports, using Qwen 2.5-14B as the teacher model and Qwen 2.5-1.5B as the student model. The proposed methodology combined hybrid distillation (combining logits and intermediate representations) with supervised fine-tuning via LoRA, aiming to preserve semantic capacity and adherence to police need while ensuring usability in environments with limited computational infrastructure.

The results of the presented approach were validated by metrics such as BERTScore-F1 and by qualitative evaluations that highlighted the clarity and narrative coherence of the summaries. The observed limitations were attributed to the generic formulation of prompts, which can be corrected with more specific instructions. The computational analysis showed expressive gains: the student model demonstrated nearly double the inference speed and a reduction of about 90% in GPU memory usage compared to the original model. Even with minor performance variations, the solution proved to be effective and robust for real-world contexts, balancing semantic precision, agility, and low operational cost.

As limitations, the absence of a standardized human reference set for the summaries stands out, which restricts the scope of the automated metrics used, such as BERTScore-F1. Furthermore, the small number of specialists involved in the qualitative evaluation could be explained by future studies for greater statistical robustness. Poorly structured or excessively short narratives also occasionally impacted the student model's performance, suggesting the need for complementary pre-processing and text normalization strategies.

For future works, it is recommended to explore the student model in other relevant textual tasks for public security, such as incident classification and automated generation of procedural documents. Additionally, comparison with other compact architectures, including Portuguese-adapted base models, can clarify whether any gains stem from base-model adaptation rather than distillation alone.

In summary, this study advances the state of the art by presenting an efficient, lightweight, and semantically robust technical solution for summarizing police reports, contributing to the strategic use of LLMs in the domain of Brazilian public security.

Acknowledgments

The authors thank the support of the Secretaria Segurança Pública do Estado de Mato Grosso (SESP-MT) through contract nº 057/2023/SESP with the Universidade Federal de Mato Grosso (UFMT).

References

Brown, T. B. et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems 33 (NeurIPS 2020), Proceedings of the 34th Conference on Neural Information Processing Systems*, pages 1877–1901.

Hinton, G., Vinyals, O., and Dean, J. (2014). Distilling the knowledge in a neural network. *Deep Learning and Representation Learning Workshop.*

Jiao, X., Yin, Y., Shang, L., Jiang, X., Chen, X., Li, L., Wang, F., and Liu, Q. (2020). Tinybert: distilling bert for natural language understanding. *Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4163–4174.

Pereira, J., Assumpção, A., Trecenti, J., Airosa, L., Lente, C., Cléto, J., Dobins, G., Nogueira, R., Mitchell, L., and Lotufo, R. (2025). Inacia: integrating large language models in brazilian audit courts: opportunities and challenges. *Digital Government: Research and Practice*, 6(1):1–20.

Ramesh, R., M, A. T. R., Reddy, H. V., and N, S. V. (2024). Fine-tuning large language models for task specific data. *Proceedings of the 2nd International Conference on Networking, Embedded and Wireless Systems (ICNEWS)*, pages 1–6.

Sarzaeim, P., Mahmoud, Q. H., and Azim, A. (2024). A framework for llm-assisted smart policing system. *IEEE Access*.

Setiawan, H. (2024). Accurate knowledge distillation via n-best reranking. *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1330–1345.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., and Fedus, W. (2022). Emergent abilities of large language models. *Transactions on Machine Learning Research*.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2020). Bertscore: evaluating text generation with bert. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Zhang, Y., Long, D., Li, Z., and Xie, P. (2023). Text representation distillation via information bottleneck principle. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14372–14383.

Zheng, X., Li, Y., Wang, H., Zhang, C., Zhang, Y., Zhou, X., Zhang, J., Zhang, J., Wang, Z., Li, K., Liu, Z., Li, L., He, X., and Wang, B. (2024). Large language models (llms): survey, technical frameworks, and future challenges. *Artificial Intelligence Review*, 58(3):1–29.