LLM4Gov: A Privacy-Preserving Approach to Teacher-Student Fine-Tuning of Distilled LLMs for the Public Sector

Ricardo M. Marcacini[1] , Jorge Carlos Valverde-Rebaza[1] , Marcelo A. S. Turine[2] , Brucce Neves Santos[1] , Silvio Levcovitz [3] , Solange O. Rezende [1]

[1]Institute of Mathematics and Computer Sciences - University of São Paulo (USP)

[2]Faculty of Computing - Federal University of Mato Grosso do Sul (UFMS)

[3]Procuradoria-Geral da Fazenda Nacional (PGFN)

{ricardo.marcacini,solange}@icmc.usp.br

{jorge.valverr,brucce.neves}@gmail.com

silvio.levcovitz@pgfn.gov.br

Abstract. *Large Language Models (LLMs) have revolutionized natural language processing, but their reliance on extensive computational resources and proprietary APIs poses significant challenges for public sector applications. Government agencies often face legal constraints, such as GDPR and LGPD, preventing the use of external LLMs when handling sensitive data. Even when compliance is met, the financial burden of deploying large-scale models remains a major barrier. To address these challenges, we present LLM4Gov, a privacy-preserving computational tool designed to fine-tune distilled LLMs for government-related tasks while minimizing infrastructure costs. LLM4Gov follows a structured teacher-student learning pipeline, where a lightweight anonymization module first removes personally identifiable information (PII) before any interaction with an external LLM. A teacher LLM then generates task-specific instructions from the anonymized dataset, and a distilled student LLM is fine-tuned using Low-Rank Adaptation (LoRA) and quantization, enabling deployment on resource-constrained environments. Our experimental results show that LLM4Gov consistently outperforms competitive distilled LLMs, achieving higher accuracy while preserving privacy and interpretability.*

1. Introduction

Good governance in the public sector involves designing innovative, evidence-based policies; ensuring transparency—especially in financial decisions; maintaining strong internal controls; and promoting accountable, ethical leadership. Today, digital transformation, driven by Artificial Intelligence and Large Language Models (LLMs), is reshaping these processes. Public policies must keep pace, promoting more inclusive, secure, and accessible services to strengthen public trust.

Large Language Models (LLMs) have transformed natural language processing, enabling advanced applications in text analysis, question answering, and automated decision-making [Chang et al. 2024]. However, their use requires high computational

resources and often depends on proprietary models like GPT-4 (OpenAI), Gemini 1.5 (Google DeepMind), Claude 3 (Anthropic), and DeepSeek. This makes them impractical due to cost and privacy concerns. The challenge is even greater for government and public sector organizations, where privacy laws and hardware limitations restrict the use of such models [Zhu et al. 2024]. Even if a government agency has the budget to pay for API access to proprietary LLMs, the recurring costs can be high, and legal regulations like LGPD in Brazil, GDPR in Europe, and other national data protection laws may prevent their direct use. These challenges have driven recent research on how to incorporate LLM advancements into the public sector while ensuring privacy and cost efficiency [Zhang et al. 2025].

A promising approach to overcoming these limitations is the use of compact distilled LLMs, which have significantly fewer parameters and can run on limited hardware while being tailored for specific tasks [Gu et al. 2023]. Several such models have recently emerged, including open versions of Llama 3.1 (Meta), Phi (Microsoft), Qwen (Alibaba), Mistral (Mistral AI), and Gemma (Google), all ranging between 7 and 14 billion parameters. These models require fewer computational resources, can run on more affordable GPUs, and can be deployed locally within an organization, reducing both cost and privacy risks. In contrast, proprietary large-scale LLMs such as GPT-4, Gemini, Claude 4, and DeepSeek are estimated to have hundreds of billions or even over a trillion parameters.

On the other hand, distilled LLMs often fail to match the performance of proprietary large-scale models, which benefit from extensive pretraining on massive datasets. To bridge this gap, Teacher-Student Fine-Tuning has gained increasing attention, where a powerful teacher LLM generates training data or supervision signals to improve the performance of a distilled student LLM [Tian et al. 2024]. This process enables the student model to capture and retain task-specific knowledge from the teacher model while significantly reducing computational requirements, preserving effectiveness in terms of response accuracy.

Despite recent advancements in Teacher-Student Fine-Tuning for LLMs, existing approaches remain inadequate for public sector applications. One major limitation is that fine-tuning a student model still requires substantial computational resources, making the process expensive and often impractical for government organizations with limited infrastructure. More critically, privacy concerns persist, as sensitive data may be exposed when using a proprietary teacher LLM to generate training data for the student LLM. This creates a significant risk, as many government and public sector datasets contain confidential or personally identifiable information that cannot be shared with external APIs or commercial models. Given these challenges, a key research question emerges: *how can we fine-tune distilled LLMs for public sector tasks while preserving data privacy and minimizing computational costs?*

In this paper, we introduce LLM4Gov, a computational tool and fine-tuning framework designed to enable the adoption of distilled LLMs in the public sector while addressing privacy and computational constraints. Our approach ensures that sensitive government data is never exposed to proprietary LLM teacher models by incorporating a privacy-preserving pipeline before fine-tuning. LLM4Gov consists of three key components: (i) an anonymization module powered by a distilled LLM to preprocess and remove personally identifiable information (PII) from training datasets, (ii) a teacher-student fine-

tuning process, where a powerful teacher LLM generates task-specific fine-tuning instructions based on the anonymized data, and (iii) a low-cost fine-tuning strategy using LoRA (Low-Rank Adaptation) and quantization [Hu et al. 2022], thereby enabling efficient training and deployment of the student LLM on resource-constrained hardware. Our main contributions can be summarized as follows:

- Privacy-Preserving Fine-Tuning Pipeline: We explore a lightweight LLaMA-based LLM that runs efficiently within the organization's local infrastructure, ensuring that sensitive data remains in a controlled environment. This anonymization module is designed to detect and remove personally identifiable information (PII), including names, dates, locations, and organizational references, before any interaction with an external teacher LLM. The module can be customized to comply with privacy regulations such as LGPD and GDPR, as well as other organization-specific policies. Additionally, we implement a sensitive information masking scheme that ensures downstream tasks rely on relevant semantic patterns rather than private data, making the model's predictions independent of sensitive attributes while preserving the usefulness of the training data.
- Task-Specific Knowledge Transfer with Anonymized Data: Instead of directly fine-tuning a distilled model on raw public sector data, LLM4Gov leverages a teacher LLM to generate instruction-based fine-tuning data from the anonymized dataset, ensuring that the generated instructions focus on the textual content and task-specific semantics rather than personal data. Additionally, LLM4Gov reinforces explanation-based learning by prompting the teacher LLM to provide rationale for each decision. This approach transfers explanatory capabilities to the student LLM while also improving the interpretability of its decisions, aligning the model with transparency requirements in public sector applications.
- Student LLM Fine-Tuning via LoRA and Quantization: To minimize computational costs, LLM4Gov adopts parameter-efficient fine-tuning (PEFT) techniques, specifically Low-Rank Adaptation (LoRA) and quantization, allowing for efficient training and deployment of LLMs in environments with limited resources. LoRA is a fine-tuning method that freezes the original model weights and adds small trainable layers, significantly reducing the number of parameters that need to be updated [Hu et al. 2022]. Quantization improves efficiency by reducing the precision of model weights and activations (e.g., from FP16 or FP32 to INT8 or lower). This lowers memory usage and speeds up inference [Jin et al. 2024].

We carried out an experimental evaluation of LLM4Gov to assess its effectiveness in real-world public sector applications. The evaluation considered legal document classification. We compared LLM4Gov against two commonly used distilled LLMs designed for privacy-aware and resource-constrained environments and the results showed that LLM4Gov outperformed existing approaches in both task accuracy and explanation quality, demonstrating its ability to provide interpretable and useful outputs while maintaining computational efficiency. We make LLM4Gov available as an open-source software in the project's GitHub repository at `https://github.com/LABIC-ICMC-USP/llm4gov`.

2. Problem Definition

Let $D = \{(x_i, y_i)\}_{i=1}^{N}$ be a training dataset, where x_i represents the input text and y_i represents the expected output for a given task. Due to privacy constraints, we define an anonymization function, as shown in Equation 1, which transforms the original dataset into an anonymized version $\tilde{D}$:

$$A : x_i \mapsto \tilde{x}_i \tag{1}$$

where $\tilde{x}_i$ is the anonymized version of x_i, ensuring that personally identifiable information (PII) and other sensitive details are removed while preserving the task-relevant semantics.

Given the anonymized dataset $\tilde{D} = \{(\tilde{x}_i, y_i)\}_{i=1}^{N}$, a teacher LLM T generates a set of task-specific instructions, as defined in Equation 2:

$$I = \{(\tilde{x}_i, s_i)\}_{i=1}^{N} \tag{2}$$

where $s_i = T(\tilde{x}_i)$ represents the instructional output generated by the teacher model T, which refines and structures the task in a format suitable for fine-tuning a student LLM.

The goal is to fine-tune a student LLM S so that its output distribution approximates that of the teacher LLM given an instruction-based prompt. We define the probability distribution of a LLM model M generating a token sequence Y given input X as follows:

$$P_M(Y|X) = \prod_{t=1}^{|Y|} P_M(y_t|X, y_{<t}) \tag{3}$$

where $P_M(y_t|X, y_{<t})$ represents the probability of the next token y_t being generated given the input X and the previously generated tokens $y_{<t}$.

To ensure that the student model S approximates the teacher model T, we aim to minimize the difference between their output distributions. This is achieved by reducing the cross-entropy loss, as defined in Equation 4:

$$\mathcal{L}(S) = -\sum_{i=1}^{N} \sum_{t=1}^{|s_i|} P_T(y_t|\tilde{x}_i, y_{<t}) \log P_S(y_t|\tilde{x}_i, y_{<t}) \tag{4}$$

where P_T represents the teacher model's probability distribution over token sequences, and P_S represents the student model's probability distribution after fine-tuning. However, the loss function in Equation 4 cannot be directly computed, as the teacher LLM does not generate full probability distributions over token sequences but rather instruction-based outputs. To approximate this objective, LLM4Gov decomposes the problem into three stages: anonymization, teacher-student instruction generation, and fine-tuning, as described in the following sections.

3. LLM4Gov

3.1. Anonymization of Sensitive Information

As described in Section 2, the first step in LLM4Gov is the transformation of the original dataset D into an anonymized version $\tilde{D}$ (Equation 1). This step ensures that personally identifiable information (PII) is removed before any interaction with the teacher LLM T, addressing privacy concerns while preserving the essential task-related semantics.

To achieve this, we employ a distilled LLaMA 3.1 model with 8 billion parameters, specifically fine-tuned for named entity recognition (NER) and anonymization tasks. The anonymization process consists of two main components:

- **Entity Recognition and Masking:** The model identifies PII entities such as names, addresses, emails, phone numbers, and organizational references, replacing them with structured placeholders (e.g., `[NAME]`, `[ADDRESS]`, `[EMAIL]`).
- **Context Preservation:** While removing sensitive information, the model ensures that sentence structure, grammar, and the core meaning of the text remain intact, allowing the downstream task to operate without relying on private attributes.

The anonymization model operates based on a structured LLM system prompt, ensuring consistency and flexibility across different types of text data. This prompt is detailed in Table 1.

System Prompt for Anonymization
"You are an expert at anonymizing sensitive information in text. Your task is to replace personal or sensitive data with generic placeholders while preserving the structure, grammar, and meaning of the original text. Ensure that you maintain the original language of the input text. The placeholders should be descriptive yet generic, such as [NAME], [EMAIL], [PHONE], [ADDRESS], [ID_NUMBER], etc. Do not alter the text beyond anonymization. The output must be a valid JSON object with exactly two keys: 'output' and 'explanation'. - 'output': The anonymized text, where sensitive entities are replaced with placeholders. - 'explanation': A brief description of the anonymized entities and their replacements."

Table 1. System prompt used for anonymization of input text.

To illustrate the anonymization process, consider the input text containing sensitive information, as shown in Table 2. After processing through the anonymization model, the output is structured as depicted in Table 3. Note that we enforce the output format to be JSON to facilitate integration with subsequent stages of the LLM4Gov pipeline.

Input Text
John Doe lives at 123 Main St, Springfield. His email is johndoe@example.com, and his phone number is (555) 123-4567.

Table 2. Example of input text before anonymization.

Anonymized Output (JSON)

```
{
    "output": "[NAME] lives at [ADDRESS]. His email is [EMAIL],
               and his phone number is [PHONE].",

    "explanation": "Replaced 'John Doe' with [NAME],
                    '123 Main St, Springfield' with [ADDRESS],
                    'johndoe@example.com' with [EMAIL],
                    and '(555) 123-4567' with [PHONE]."
}
```

Table 3. Example of anonymized output in JSON format.

3.2. Task-Specific Knowledge Transfer with Anonymized Data

As defined in Section 2, LLM4Gov aims to fine-tune a student LLM S to approximate the behavior of a larger teacher LLM T, while ensuring privacy by operating on the anonymized dataset $\tilde{D}$ obtained through Equation 1. Since the direct computation of the loss function in Equation 4 is not feasible due to the nature of instruction-based outputs, LLM4Gov approximates this process by generating structured fine-tuning data using a teacher LLM.

For each anonymized input $\tilde{x}_i$, the teacher LLM T generates an instruction s_i according to Equation 2. Remeber that the goal is to define a training set in the form of structured instruction-based data $I = \{(\tilde{x}_i, s_i)\}_{i=1}^{N}$, where s_i is not a direct token sequence distribution but rather a structured response that provides task-specific guidance for fine-tuning the student model. The structured format follows a triplet representation:

$$\langle \text{instruction}, \text{input}, \text{output} \rangle \tag{5}$$

where instruction specifies the task to be performed, input $= \tilde{x}_i$ is the anonymized document, and output is a structured response in JSON format, containing model-generated predictions and explanations. If labeled data exists, the JSON output is pre-filled with annotations, guiding the teacher LLM to focus on explaining the labeled information rather than inferring it.

The instruction generation follows a structured system prompt that standardizes task definition, input handling, and expected output formatting. Table 4 provides an overview of this prompt.

System Prompt Structure for Instruction Generation

Task Definition: instruction
Input Document (Anonymized): anonymized_text
Expected Output Schema (JSON Format): json_output

Table 4. General structure of the system prompt for instruction generation.

The generated dataset I forms the basis for fine-tuning the student LLM S, ensuring that it learns task-specific knowledge while aligning its probability distribution $P_S(Y|\tilde{x}_i)$ with the teacher's probability distribution $P_T(Y|\tilde{x}_i)$, as described in Equation 4.

3.3. Student LLM Fine-Tuning via LoRA and Quantization

LLM4Gov fine-tunes a LLaMA 3.1 model with 8 billion parameters using 4-bit quantization, significantly reducing memory requirements. The quantization process converts high-precision floating-point model weights into lower-bit representations. The final weight representation after quantization is given by

$$W_q = Q(W_0 + AB) \tag{6}$$

where $Q(\cdot)$ is the quantization function that maps floating-point values to lower-bit representations (e.g., INT4). This allows LLM4Gov to run on lower-end GPUs while maintaining fine-tuning flexibility through LoRA updates.

Fine-tuning a full LLM requires updating all model parameters, which is computationally expensive. Instead, LLM4Gov use LoRA to freeze the pre-trained model weights and introduces trainable low-rank matrices, reducing the number of updated parameters while preserving the knowledge captured during pretraining. Let W_0 be the original weight matrix of a given layer. Instead of directly updating W_0, LoRA introduces two low-rank matrices $A \in \mathbb{R}^{d\times r}$ and $B \in \mathbb{R}^{r\times k}$, where r is the rank of the update. The adapted weight matrix is then computed as

$$W = W_0 + AB \tag{7}$$

where W_0 remains frozen and only the low-rank matrices A and B are trained. Since $r \ll d, k$, this significantly reduces the number of trainable parameters.

Using the notation from Section 2, let $P_S(y_t|\tilde{x}_i, y_{<t})$ represent the probability distribution of the student LLM over token sequences, as defined in Equation 3. The objective of fine-tuning in LLM4Gov is to update the student model's parameters to minimize the loss function in Equation 4 while keeping most of the original model weights frozen. By incorporating the LoRA update from Equation 7, the probability of generating a token y_t is redefined as

$$P_S(y_t|\tilde{x}_i, y_{<t}, W) = P_S(y_t|\tilde{x}_i, y_{<t}, W_0 + AB) \tag{8}$$

Since only A and B are trainable, the fine-tuning optimization is constrained to

$$\mathcal{L}(A, B) = -\sum_{i=1}^{N}\sum_{t=1}^{|s_i|} P_T(y_t|\tilde{x}_i, y_{<t}) \log P_S(y_t|\tilde{x}_i, y_{<t}, W_0 + AB) \tag{9}$$

where the objective remains to align P_S with P_T, but now using a parameter-efficient update mechanism.

4. Experimental Evaluation

4.1. Datasets

We used a legal document dataset, corresponding to a specific task relevant to public sector applications. The texts were preprocessed by the anonymization module of LLM4Gov before any further processing. Table 5 summarizes the main characteristics of each dataset, including the number of training and test instances and the targets generated by the teacher LLM.

Dataset	Instances (Train/Test)	Task	Targets (JSON Keys)
Legal Documents	1984 / 851	Classification	`"category"`

Table 5. Characteristics of the datasets used for evaluation.

An anonymized version of the dataset used in our experiments will be made publicly available in the final version of the paper to ensure the reproducibility of our results. Alongside the dataset, we will also release the source code of the LLM4Gov framework and all scripts used for training and evaluation.

4.2. Evaluation Criteria

We compared LLM4Gov against two commonly used distilled LLMs designed for privacy-aware and resource-constrained environments: LLaMA 3.1 (8B) and DeepSeek-R1 (7b). All models were evaluated using the same prompts in a zero-shot setting, ensuring a fair comparison. The comparison with LLaMA 3.1 is particularly relevant, as it served as the base model for fine-tuning LLM4Gov. DeepSeek-R1, on the other hand, represents a recent approach that applies reasoning-oriented fine-tuning to LLaMA models, offering an interesting benchmark for evaluating instruction-following and reasoning capabilities in compact LLMs.

For each baseline, the anonymized test data was used as input, and the models were prompted to produce outputs in the same structured JSON format as LLM4Gov. In LLM4Gov, the anonymized data was also used during inference, ensuring consistency across evaluations.

To measure performance, we adopted Precision, Recall, and F1-score, computed using BertScore [Zhang et al. 2019]. This approach evaluates the similarity between the outputs of the distilled models and those of the teacher LLM, assessing both correctness and semantic closeness. Given that outputs are structured JSON objects, the evaluation was performed on a per-target basis, comparing each generated key individually.

4.3. Results

Table 6 presents the evaluation results. The results indicate that LLM4Gov consistently outperformed all baselines in terms of Precision, Recall, and F1-score, demonstrating its ability to generate outputs more aligned with the teacher LLM while maintaining computational efficiency.

Across all datasets, LLM4Gov achieved the highest F1-score, confirming that its fine-tuning strategy enhances the quality of outputs compared to other distilled LLMs.

Model	Precision (P)	Recall (R)	F1
LLaMA 3.1 (8B)	0.765	0.748	0.756
DeepSeek-R1 (7B)	0.663	0.698	0.680
LLM4Gov (ours)	**0.788**	**0.789**	**0.788**

Table 6. Evaluation results for the legal document classification dataset.

The performance improvements suggest that the teacher-student training pipeline effectively transfers knowledge while preserving privacy. The consistent gains in precision and recall also indicate that the method produces more reliable and interpretable outputs, making it well-suited for government applications. Despite being fine-tuned for reasoning tasks, DeepSeek-R1 did not perform as well as expected, achieving lower scores across all metrics. This suggests that reasoning capabilities alone are not sufficient for domain-specific tasks such as legal text classification. The lack of fine-tuning with in-domain data appears to limit its effectiveness, reinforcing the importance of aligning models with the linguistic and conceptual nuances of the target domain.

To illustrate the evaluation metrics in practice, the following comparison provides a concrete example contrasting the Ground Truth annotations with the output generated by the fine-tuned `LLM4Gov` model and the predictions from the base `LLaMA 3.1` model (without fine-tuning). Each output includes four fields—summary (`resumo`), keywords (`palavras_chave`), labels (`rotulos`), and explanation (`explicacao`) — enabling a analysis of performance and alignment with the original document.

Ground Truth


```
{
  "resumo": "O Estado de [STATE] solicitou a suspensão de todos os
     processos nacionais relacionados    mesma questão jurídica
     discutida no recurso extraordinário, que trata do direito dos
     servidores públicos a indenização devido    falta de reposição do
      poder aquisitivo dos vencimentos. O Ministro relator indeferiu o
      pedido, apontando a excepcionalidade da suspensão e o possível
     conflito com o acesso ao Judiciário.",
  "palavras_chave": ["suspensão de processos", "direito dos servidores"
     , "indenização", "reposição do poder aquisitivo", "repercussão
     geral"],
  "rotulos": ["SISTEMA REMUN", "SERVIDOR PUBLICO CIVIL", "DIREITO
     ADMINISTRATIVO E OUTRAS MATERIAS DE DIREITO PUBLICO"],
  "explicacao": "Os rótulos foram atribuídos com base em partes especí
     ficas do texto: 'SISTEMA REMUN' está relacionado ao tema da
     reposição do poder aquisitivo dos vencimentos dos servidores,
     indicando preocupação com as condições remuneratórias. 'SERVIDOR
     PUBLICO CIVIL' é relevante porque o texto discute o direito dos
     servidores a uma indenização. 'DIREITO ADMINISTRATIVO E OUTRAS
     MATERIAS DE DIREITO PUBLICO' estão presentes, pois o caso envolve
      uma questão jurídica de ordem administrativa, que afeta a
     administração pública e o direito dos servidores."
}
```


LLM4Gov (Fine-Tuned)

```
{
  "resumo": "O texto trata de um pedido de suspensão de processos
     judiciais pendentes sobre a matéria do direito a indenização por
     inobservância da reposição do poder aquisitivo dos vencimentos,
     com repercussão geral reconhecida pelo Supremo Tribunal Federal.
     O pedido de suspensão foi indeferido, pois a norma processual
     questionada seria excepcional, levando em consideração o tempo
     necessário para julgar todos os processos e a garantia
     constitucional de acesso ao Judiciário.",
  "palavras_chave": ["repercussão geral", "indenização", "direito
     aquisitivo", "Código de Processo Civil", "jurisdição"],
  "rotulos": ["DIREITO DO TRABALHO", "SISTEMA REMUNERATORIO E BENE", "
     BASE DE CALCULO", "JURISDICAO E COMPETENCIA"],
  "explicacao": "O texto aborda a suspensão de processos, jurisdição e
     a discussão de matéria jurídica relevante, que inclui a reposição
      do poder aquisitivo dos vencimentos, o que está relacionado a
     direito trabalhista e sistema remuneratório, bem como a base de c
     álculo de benefícios e o direito a indenização. Além disso, a men
     ção ao Código de Processo Civil e ao direito de acesso ao Judiciá
     rio reforça a relação com o direito administrativo e matéria
     processual civil."
}
```


LLaMA 3.1 (Base Model)


```
{
  "resumo": "Petição ao STF para determinar a suspensão de processos
     pendentes sobre a definição do direito dos servidores a indenizaç
     ão, ante a inobservância da cláusula constitucional da reposição
     do poder aquisitivo dos vencimentos.",
  "palavras_chave": ["Repercussão Geral", "Sobrestamento de Processos",
      "Direitos dos Servidores", "Indenização"],
  "rotulos": ["DIREITO DO TRABALHO", "LIMITACOES AO PODER DE TRIBUTAR",
      "EXTINCAO DO CREDITO TRIBUTARIO"],
  "explicacao": "Os rótulos foram escolhidos com base na discussão
     sobre os direitos dos servidores, indenização e limitações ao
     poder de tributar. A petição visa determinar a suspensão de
     processos pendentes sobre essa matéria, o que está relacionado
      extinção do crédito tributário."
}
```


LLM4Gov demonstrates a better alignment with the Ground Truth by correctly identifying core legal themes such as ”SISTEMA REMUNERATORIO”, ”JURISDICAO”, and the broader administrative and labor law context. Its summary and explanation show a more nuanced understanding of the legal reasoning and structure compared to the base model. In contrast, LLaMA 3.1 (Base) captures the general idea but misclassifies the domain-specific legal labels, introducing unrelated categories such as ”LIMITACOES AO PODER DE TRIBUTAR”, which are not contextually appropriate.

5. Related Work

The growing need for efficient adaptation of Large Language Models (LLMs) has led to the development of teacher-student fine-tuning techniques, particularly for reducing computational costs while preserving performance. Several recent works have explored

knowledge distillation approaches, including TinyLLM [Tian et al. 2024] and MiniLLM [Gu et al. 2023]. These methods have demonstrated significant improvements in reasoning and efficiency by transferring structured knowledge from large teacher models to compact student models.

Despite these advances, existing works do not fully address the challenges of applying teacher-student fine-tuning in the public sector, where privacy, compliance, and deployment constraints are critical. The literature lacks methods specifically tailored for governmental AI applications, where models must be fine-tuned on anonymized data while maintaining transparency, accountability, and interpretability. Table 7 compares existing approache, highlighting gaps in the current literature.

Approach	**Key Idea**	**Advantages**	**Limitations**
TinyLLM [Tian et al. 2024]	Multi-teacher distillation with Chain-of-Thought reasoning.	Improves student model reasoning and accuracy. Outperforms larger LLMs in certain benchmarks.	Requires multiple LLM teachers, increasing computational cost. Focused on reasoning, not general NLP.
MiniLLM [Gu et al. 2023]	KL divergence optimization for better knowledge transfer.	Enhances response accuracy and calibration. Generalizable across multiple LLM families.	Assumes white-box access to teacher models. No focus on privacy.
LLM4Gov (Proposed)	Teacher-student fine-tuning with anonymized training data for public sector applications.	Domain-specialized model for government tasks. Ensures privacy compliance. Enables efficient deployment in local infrastructures.	Reliance on teacher-generated instructions. Requires further optimization for teacher LLM bias mitigation.

Table 7. Comparison of existing fine-tuning approaches and LLM4Gov.

We do not perform a direct empirical comparison with TinyLLM or MiniLLM, as these approaches were not designed with privacy-aware constraints in mind. Both methods assume unrestricted access to training data and teacher models, without addressing the strict requirements of anonymization, compliance, and local deployment that are essential in public sector scenarios.

6. Concluding Remarks

In this work, we introduced LLM4Gov, a privacy-preserving teacher-student fine-tuning framework designed for public sector applications. Our approach ensures compliance with data protection regulations by employing an anonymization module that removes personally identifiable information (PII) before interacting with an external teacher LLM. LLM4Gov further enhances model efficiency by using instruction-based fine-tuning on anonymized data and leveraging parameter-efficient fine-tuning techniques, such as LoRA and quantization, to reduce computational costs.

Through experimental evaluation on three real-world government-related tasks—legal document classification, electronic invoice enrichment, and research project categorization—LLM4Gov consistently outperformed competitive distilled LLMs across all datasets. The results demonstrated that our approach not only achieves higher accuracy in task resolution but also generates more interpretable outputs due to the integration of explanation-based learning during the fine-tuning process. Despite its advantages, LLM4Gov presents some limitations. First, while anonymization mitigates privacy risks, it may inadvertently remove useful contextual information, potentially impacting model performance in specific scenarios. Second, the reliance on teacher-generated instructions

means that the quality of fine-tuning data depends on the robustness of the teacher LLM, which may introduce biases or inconsistencies.

Future work includes exploring adaptive anonymization strategies that preserve task-relevant context while ensuring privacy. Additionally, we aim to investigate more efficient knowledge transfer techniques to further improve the generalization of student models. Another direction involves integrating LLM4Gov into public sector platforms, enabling real-world deployment and evaluation in large-scale government AI applications. For example, integrating LLM4Gov into the "Observatory of the National Agenda for Graduate Education" of CAPES/MEC aims to support the governance of public policy focused on the training and retention of master's and doctoral degree holders in the strategic and priority areas identified by Brazil's 27 federal units.

Acknowledgements: This work was supported by the National Council for Scientific and Technological Development (CNPq), Grant No. 316507/2023-7, the São Paulo Research Foundation (FAPESP), Grant No. 2023/10100-4, and the Coordination of Superior Level Staff Improvement (CAPES), Grant No. 14984/2024.

References

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., et al. (2024). A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3):1–45.

Gu, Y., Dong, L., Wei, F., and Huang, M. (2023). Minillm: Knowledge distillation of large language models. In *The Twelfth International Conference on Learning Representations*.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al. (2022). Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Jin, R., Du, J., Huang, W., Liu, W., Luan, J., Wang, B., and Xiong, D. (2024). A comprehensive evaluation of quantization strategies for large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 12186–12215.

Tian, Y., Han, Y., Chen, X., Wang, W., and Chawla, N. V. (2024). Tinyllm: Learning a small student from multiple large language models. *arXiv e-prints*, pages arXiv–2402.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2019). Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Zhang, X., Pang, Y., Kang, Y., Chen, W., Fan, L., Jin, H., and Yang, Q. (2025). No free lunch theorem for privacy-preserving llm inference. *Artificial Intelligence*, page 104293.

Zhu, X., Li, J., Liu, Y., Ma, C., and Wang, W. (2024). A survey on model compression for large language models. *Transactions of the Association for Computational Linguistics*, 12:1556–1577.