RECIMA21 - REVISTA CIENTÍFICA MULTIDISCIPLINAR
ISSN 2675-6218

TRANSFORMING THE LEGAL LANDSCAPE: AN AI-DRIVEN FRAMEWORK FOR JUDICIAL TEXT PROCESSING

TRANSFORMANDO O CENÁRIO JURÍDICO: UMA ESTRUTURA ORIENTADA POR IA PARA PROCESSAMENTO DE TEXTOS JUDICIAIS

TRANSFORMANDO EL PANORAMA LEGAL: UNA ESTRUCTURA IMPULSADA POR IA PARA EL PROCESAMIENTO DE TEXTOS JUDICIALES

Luciano Zanuz[1], Sandro José Rigo[2]

e656426
https://doi.org/10.47820/recima21.v6i5.6426
RECEIVED: 03/16/2025 APPROVED: 04/16/2025 PUBLISHED: 05/01/2025

ABSTRACT
Artificial Intelligence can revolutionize the legal field by addressing the complexities of managing extensive textual data inherent to judicial processes. Nevertheless, the literature highlights the difficulties in managing different contexts regarding distinct application scenarios. This paper presents a novel methodology tailored for developing applications in the legal domain, leveraging cutting-edge natural language processing techniques, including transformer-based architecture, pre-trained models, and transfer learning. Unlike traditional software development, this approach embraces the inherent uncertainties of Artificial Intelligence solutions by employing an iterative framework that integrates strong collaboration with legal professionals, domain-specific datasets, and comprehensive evaluation strategies. The methodology was validated through real-world applications at the Court of Justice of Rio Grande do Sul, including the development of a Judgment Report Generator, which automates judgment report creation using Generative Artificial Intelligence, and additional experiments showcased state-of-the-art performance in legal Named Entity Recognition using fine-tuned BERT models and context-adapted text generation with GPT-2-based models, demonstrating adaptability to diverse legal scenarios. This work bridges advanced natural language processing techniques with the practical demands of the judiciary, establishing a foundation for scalable, reliable, and domain-aware AI applications. The proposed methodology addresses practical challenges, regulatory alignment, and dataset specificity, enabling effective AI integration in the legal sector for enhanced efficiency and impact in real-world judicial systems.

KEYWORDS: Artificial Intelligence and Law. Natural Language Processing. Methodology for Artificial Intelligence Development.

RESUMO
A Inteligência Artificial pode revolucionar o campo jurídico ao abordar as complexidades do gerenciamento de extensos dados textuais inerentes aos processos judiciais. No entanto, a literatura destaca as dificuldades em gerenciar diferentes contextos em relação a distintos cenários de aplicação. Este artigo apresenta uma nova metodologia adaptada para o desenvolvimento de aplicações no domínio jurídico, alavancando técnicas de processamento de linguagem natural de ponta, incluindo arquiteturas baseadas em transformadores, modelos pré-treinados e aprendizado por transferência. Diferentemente do desenvolvimento de software tradicional, essa abordagem abrange as incertezas inerentes às soluções de Inteligência Artificial, empregando uma estrutura iterativa que integra forte colaboração com profissionais do direito, conjuntos de dados específicos do domínio e estratégias abrangentes de avaliação. A metodologia foi validada por meio de aplicações reais no Tribunal de Justiça do Rio Grande do Sul, incluindo o desenvolvimento de um Gerador de Relatórios de Julgamento, que automatiza a criação de relatórios de julgamento usando Inteligência Artificial Generativa, e experimentos adicionais demonstraram desempenho de ponta em Reconhecimento de

[1] PhD student in Artificial Intelligence applied to Law at UNISINOS. MsC (2009) and Bsc (2000) in Applied Computing at UNISINOS and specialization at UFRGS (2004). Systems analyst at the Rio Grande do Sul State Court of Justice and professor at UNISENAC-RS.
[2] Bsc in Computer Science at PUCRS (1990); MsC (1993) and PhD (2008) in Computer Science at UFRGS (2008). Post-doctorate at Friedrich-Alexander Universität Erlangen-Nürnberg/Germany (2018). Professor at UNISINOS; Researcher in UNISINOS/PPGCA. Dean of the UNISINOS Polytechnic School.

RECIMA21 - REVISTA CIENTÍFICA MULTIDISCIPLINAR
ISSN 2675-6218

Entidades Nomeadas jurídicas usando modelos BERT ajustados e geração de texto adaptada ao contexto com modelos baseados em GPT-2, demonstrando adaptabilidade a diversos cenários jurídicos. Este trabalho conecta técnicas avançadas de processamento de linguagem natural com demandas práticas do judiciário, estabelecendo uma base para aplicações de IA escaláveis e confiáveis. A metodologia proposta aborda desafios práticos, alinhamento regulatório e especificidade do conjunto de dados, permitindo a integração eficaz da IA no setor jurídico para maior eficiência e impacto em sistemas judiciais do mundo real.

PALAVRAS-CHAVE: Inteligência Artificial e Direito. Processamento de Linguagem Natural. Metodologia para Desenvolvimento de Inteligência Artificial.

RESUMEN
La Inteligencia Artificial puede revolucionar el ámbito jurídico al gestionar grandes volúmenes de datos textuales en los procesos judiciales. Sin embargo, existen desafíos para manejar diferentes contextos en diversos escenarios. Este artículo propone una metodología innovadora para desarrollar aplicaciones jurídicas, aprovechando técnicas avanzadas de procesamiento del lenguaje natural, como arquitecturas basadas en transformadores, modelos preentrenados y aprendizaje por transferencia. A diferencia del desarrollo tradicional de software, este enfoque aborda las incertidumbres propias de la IA con un marco iterativo que integra colaboración con profesionales del derecho, datos específicos del dominio y estrategias de evaluación integrales. La metodología fue validada con aplicaciones reales en el Tribunal de Justicia de Rio Grande do Sul, incluyendo un Generador de Informes de Sentencia basado en IA Generativa. Además, se obtuvieron resultados de vanguardia en el Reconocimiento de Entidades Nombradas con modelos BERT optimizados y generación de texto contextual con modelos basados en GPT-2, mostrando su adaptabilidad a distintos contextos legales. Este trabajo conecta técnicas avanzadas de lenguaje natural con las necesidades del poder judicial, sentando las bases para aplicaciones de IA escalables y fiables. La metodología propuesta enfrenta desafíos prácticos, armonización regulatoria y especificidad de los datos, permitiendo una integración eficaz de la IA en el sector legal y mejorando la eficiencia e impacto de los sistemas judiciales reales.

PALABRAS CLAVE: Inteligencia Artificial y Derecho. Procesamiento del Lenguaje Natural. Metodología para el Desarrollo de Inteligencia Artificial.

INTRODUCTION

Artificial Intelligence (AI) has emerged as a transformative force in various domains, including the legal field, where it holds the potential to revolutionize how legal professionals process, analyze, and utilize information.

The legal sector is characterized by the need to manage vast volumes of textual data, ranging from case law and statutes to contracts and legal opinions. Traditional methods of handling such data often involve manual, time-intensive tasks prone to errors and inconsistencies. AI-driven applications, particularly those leveraging natural language processing (NLP), are being developed to address these challenges, enabling faster, more accurate, and cost-effective solutions for tasks like legal research, automated document drafting, summarization, and entity recognition, among many other legal processes that can benefit from AI-driven automation and analysis.

Recent advances in NLP have been pivotal in enhancing the effectiveness of AI applications in the legal domain. State-of-the-art methodologies such as transformer-based architectures (Vaswani *et al.,* 2017), pre-trained language models (PLMs) like BERT (Devlin *et al.,* 2019) and GPT (Radford *et al.,* 2018), and techniques grounded in transfer learning (Pan; Yang, 2010) have set new

RECIMA21 - REVISTA CIENTÍFICA MULTIDISCIPLINAR
ISSN 2675-6218

benchmarks in the ability of AI systems to understand and process complex textual data. These advancements allow AI systems to capture nuanced meanings, contextual relationships, and domain-specific terminology essential for legal applications. The ability to fine-tune pre-trained models for specific legal tasks further enables the development of customized solutions tailored to meet the unique demands of the field.

However, the development of AI applications diverges significantly from traditional software development paradigms. Traditional applications are typically built with a clear understanding of the problem space, allowing the development team to design a technical solution with predictable outcomes. In contrast, AI-driven applications, particularly in the domain of NLP, present inherent uncertainties. At the outset of a project, it is often unclear what model architecture, training data, or optimization strategies will yield the best results. Additionally, there is no guarantee that the solution will achieve the level of accuracy required by the business objectives. This uncertainty necessitates an iterative, experimental approach to AI development, where solutions evolve through continuous testing, evaluation, and refinement.

This article introduces a novel methodology for developing AI applications in the legal domain by leveraging the latest advancements in transformers, pre-trained models, and transfer learning. The proposed approach addresses the unique challenges of AI development, including the uncertainties in solution predictability and performance, by providing a structured framework for optimizing legal NLP solutions. By bridging cutting-edge NLP techniques with the practical requirements of legal applications, this methodology aims to advance the integration of AI into the legal field, ultimately enhancing efficiency, accuracy, and accessibility in legal services.

The following sections outline the methodology, detail its technical aspects and components, present the results and contributions, and conclude with a discussion on the proposed approach.

1. METHODOLOGY

This work introduces a tailored methodology for developing AI applications in the legal domain, leveraging state-of-the-art natural language processing technologies, including transformer-based architectures, pre-trained models, and transfer learning. By combining these advanced techniques, the methodology provides a robust framework for creating scalable and highly effective AI solutions tailored to the unique needs of the legal field. While designed to be adaptable across different languages and sectors, this research focuses on the legal context and the Portuguese language, addressing their specific challenges and complexities.

Unlike traditional application development, where requirements, outcomes, and technical solutions are typically well-defined and predictable, AI applications introduce a significant layer of uncertainty. It is often unclear at the outset which technical approach will yield the desired accuracy or whether it can satisfy the business requirements. This unpredictability highlights the distinct nature of AI projects compared to conventional software development.

The key differences stem from two critical factors:

RECIMA21 - REVISTA CIENTÍFICA MULTIDISCIPLINAR
ISSN 2675-6218

1. Data dependence. Traditional applications rely on predetermined business rules and requirements, allowing developers to create comprehensive test cases that ensure functionality. In contrast, AI applications are inherently data-driven. The system's performance depends on the data it processes, which cannot always be exhaustively mapped or anticipated during development.
2. Uncertainty in technical capabilities. Traditional development teams typically work with established technologies, tools, and frameworks whose capabilities and limitations are well understood. Conversely, AI development often involves exploring new, untested technologies, requiring the team to approach the project as a research and development endeavor. This necessitates experimentation and iterative refinement to confirm the technical feasibility of the solution.

To address these challenges, we propose a methodology that incorporates the following strategies to enhance the success of AI applications in the legal domain:

1. Strong collaboration with the business area. Engaging closely with legal professionals ensures that the solution aligns with domain-specific requirements and maximizes its utility and adoption.
2. Creation of application-specific datasets. Developing datasets tailored to the legal domain enables robust validation of the AI solution, ensuring its relevance and performance in real-world scenarios.
3. Well-defined evaluation frameworks. Establishing clear criteria and metrics for evaluating experiments provides a structured approach to measuring progress and iterating on the solution.
4. Incremental development. Building and testing preliminary versions, such as Proof of Concept (POC) and Minimum Viable Product (MVP), allows the team to assess the solution's quality, validate technical assumptions, and refine its adherence to business requirements.

This methodology has been shaped by extensive experimentation and the practical experience gained from developing real-world AI applications at the Court of Justice of Rio Grande do Sul, in southern Brazil. It underscores the importance of combining technical innovation with active participation from the business domain to ensure project success.

Figure 1 provides a schematic overview of the proposed methodology, with further details explored in subsequent sections. Central to this approach is the active collaboration with legal professionals, ensuring that the AI solution is closely aligned with real-world requirements and challenges. By integrating transformers, pre-trained models, and transfer learning, this methodology not only enables the development of effective and scalable AI solutions but also emphasizes the critical role of business domain expertise in shaping solutions that are both practical and impactful. This focus on collaboration sets the framework apart from traditional development models, ensuring that the technology serves the specific needs of the legal field.

RECIMA21 - REVISTA CIENTÍFICA MULTIDISCIPLINAR
ISSN 2675-6218

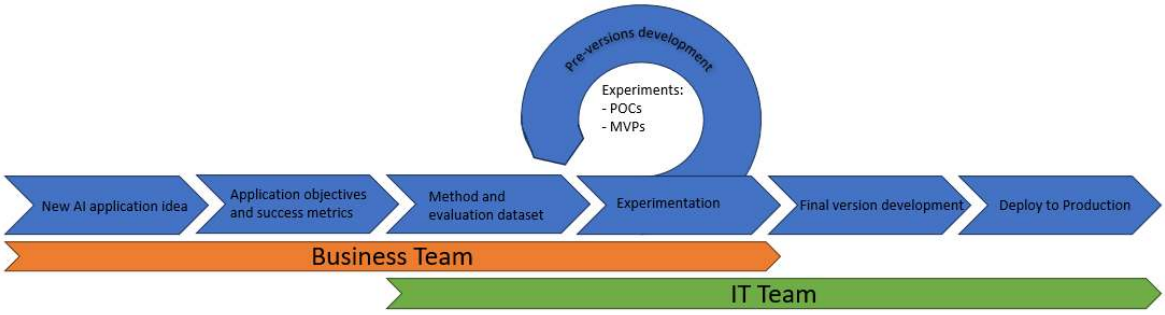

Figure 1. Overview of the methodology

1.1. New AI Application Idea

The methodology begins with the initiation of a new AI application, which has already undergone the organization's internal innovation and AI demand processes. These ideas have been vetted and approved by relevant authorities, such as committees or working groups composed of Product Owners (POs) and key stakeholders, making them ready for development. This stage is primarily the responsibility of the Business Team, and it is outside the scope of this work to define or address the approval process for AI application ideas.

1.2. Application Objectives And Success Metrics

Following the approval of the AI application idea, our methodology emphasizes the need to define clear application objectives and success metrics. At this stage, the Business Team is responsible for articulating the problem to be solved or identifying the opportunity for improvement.

In addition to clarifying the problem, this stage also involves the Business Team setting specific objectives to be achieved and defining the criteria for measuring the success of the project.

It is important to note that, up to this point, the IT Team's involvement is minimal, typically limited to providing occasional clarification on technical aspects. A crucial aspect of our methodology is the strong commitment expected from the Business Team, which goes beyond the usual level of involvement seen in traditional projects. In many cases, the business area may request a solution but fails to engage deeply throughout the development process. While this issue can arise in conventional projects, it has a particularly significant impact in AI initiatives, where close collaboration from the business side is essential for success.

1.3. Method and Evaluation Dataset

As discussed in section AI and Law application, the success of an AI application extends beyond simply selecting a well-trained model. It involves specific development tailored to the problem at hand and the business requirements. By a "well-trained AI model", we refer to a model that performs well on standard benchmark datasets, such as MMLU (Hendrycks *et al.,* 2021), HellaSwag (Zellers *et al.,* 2019), and the AI2 Reasoning Challenge (ARC) (Clark *et al.,* 2018), for example.

RECIMA21 - REVISTA CIENTÍFICA MULTIDISCIPLINAR
ISSN 2675-6218

In the context of legal AI applications, the end user typically seeks automation in judicial processes or tasks such as automatic prediction, analysis, classification, or information extraction applied to legal proceedings. AI often serves as the technological backbone for these tasks. To ensure the solution's effectiveness, it is crucial that evaluation is conducted using application-specific datasets, derived from the organization's own data and aligned with the defined application objectives. These datasets, which we refer to as application datasets, are essential for assessing the model's real-world performance. For instance, to evaluate an application designed to classify judicial processes into specific STJ or STF themes, the evaluation must involve a dataset containing documents and corresponding themes for classification. Similarly, applications that classify court cases by judicial subjects or classes would require corresponding datasets.

There are notable differences between benchmarking datasets and application datasets:

1. Purpose of use
    a. Benchmarking dataset is used to evaluate and compare the performance of models on specific tasks, serving as a standard reference to test models under controlled conditions.
    b. Application dataset is used to assess a model in a specific real-world application context. It reflects the actual data that the model will encounter in production and is tailored to the needs of the problem it aims to solve.
2. Generality versus specificity
    a. Benchmarking dataset is typically composed of standardized, widely used data within the research community. These datasets cover a range of scenarios, ensuring that models are robust and capable of generalizing across various contexts.
    b. Application dataset is specific to the application domain, containing particularities of the real-world data that the model will process. These datasets tend to be more representative of the operational environment where the model will be deployed.

The development of the application dataset is a collaborative effort between the Business and IT teams. The Business Team contributes by selecting and defining the data, while the IT Team creates the necessary scripts to generate the dataset. From this point onward, the IT Team becomes directly involved in the project, marking a shift from the initial business-focused phase.

For each AI application, an evaluation method must be defined alongside the dataset. Although the use of scientific methods is encouraged, it is not always mandatory, particularly when the application is unlikely to be published, and the timeline for delivery is short. The primary goal is to ensure that the evaluation effectively measures the solution's quality and its applicability to real-world scenarios.

Whenever possible, standard automatic metrics like accuracy, precision, ROUGE (Lin, 2004), BLEU (Papineni *et al.,* 2002), and others should be used. In cases where automatic metrics fall short

RECIMA21 - REVISTA CIENTÍFICA MULTIDISCIPLINAR
ISSN 2675-6218

of confirming the solution's quality, alternative strategies such as LLM-as-a-judge (Zheng *et al.,* 2023) or human assessment can be considered.

When selecting the final solution, it is crucial that the evaluation not only consider the metrics but also the cost-benefit ratio. For instance, even if a model like GPT-4 demonstrates superior metrics, the team may opt for a simpler version (e.g., GPT-4-mini) if the latter offers comparable results at a lower cost, including token usage.

Additionally, processing time is another critical factor. In scenarios where the application requires real-time responses, the project team may need to balance quality with speed, opting for solutions that meet the real-time constraints even if they sacrifice some performance.

1.4. Experimentation

In contrast to traditional application development, where the technical feasibility of a solution can generally be guaranteed, AI application development often involves a degree of uncertainty. To address this, we propose an experimentation phase that includes prototyping rapid solutions, commonly known as Proof of Concepts (POCs) or Minimum Viable Products (MVPs). This stage allows for the evaluation of the solution before its final development.

During this phase, the IT Team plays a key role in developing preliminary versions of the application, such as POCs or MVPs, to assess the quality of the solution and its alignment with business requirements. The Business Team is involved as the evaluator of these prototypes, providing feedback on whether they meet the business objectives.

The experiments in this stage do not need to result in fully functional applications. Instead, they may consist of reports, spreadsheets, or other outputs generated by procedures or scripts. In cases where a simple prototype application is needed, we recommend using market tools like Streamlit[1] and Gradio[2], or support tools such as Jupyter notebooks[3] and Google Colaboratory[4].

Each experiment must be evaluated using the method and evaluation dataset established in the previous stage to determine whether it meets the business area's expectations. If the solution does not meet the required criteria, the IT Team will create a new prototype, experimenting with different technological approaches in an effort to achieve the defined success metrics. Throughout this phase, the Business Team remains actively involved, assessing the prototypes and deciding whether they meet the requirements for approval to proceed to the final development and deployment stages.

At the conclusion of the experimentation phase, a validated technical solution is in place, ready to be refined and scaled for production deployment.

[1] https://streamlit.io/
[2] https://www.gradio.app/
[3] https://jupyter.org/
[4] https://colab.research.google.com/

RECIMA21 - REVISTA CIENTÍFICA MULTIDISCIPLINAR
ISSN 2675-6218

1.5. Final Version Development

At this stage, the IT Team develops a fully functional application based on the technical solution established in the approved pre-version. This phase typically involves integrating technical requirements such as security, performance, scalability, and logging, among others. While the Business Team is generally not involved at this stage, their participation may be required occasionally to address business-related questions, such as how the application will be incorporated into the broader legal proceedings system.

The final version of the application must comply with Resolution No. 332 issued by the National Council of Justice[5], which outlines specific guidelines for constructing artificial intelligence systems for the Brazilian judiciary. In accordance with the resolution's provisions on "ethics, transparency, and governance", all inferences made by the AI model, including inputs and outputs, must be logged for future analysis and auditing. Additionally, it is recommended to log processing times, token usage, and associated costs to facilitate a comprehensive cost-benefit analysis of the solution.

To ensure adherence to the resolution, if the final application relies on an internally trained PLM or LLM, as further discussed in section 2, the organization must maintain copies of the datasets used for training, along with the corresponding model versions. This ensures that the system can be analyzed and audited in the future, should the need arise.

1.6. Deploy to Production

The final stage of the methodology is primarily technical and falls under the responsibility of the IT Team. Its main objective is to deploy the final version of the application to the production environment, enabling users to begin using it and achieve the desired results as defined at the outset of the project.

2. TECHNICAL FRAMEWORK FOR AI AND LAW APPLICATIONS DEVELOPMENT

In this section, we describe the key technical aspects of the proposed methodology that the IT Team must consider when developing experiments and the final version of the AI and Law application. ***Figure 2*** provides an overview of these technical components, which are divided into four distinct layers. Each layer is further explained in the subsections that follow.

[5] https://atos.cnj.jus.br/atos/detalhar/3429

RECIMA21 - REVISTA CIENTÍFICA MULTIDISCIPLINAR
ISSN 2675-6218

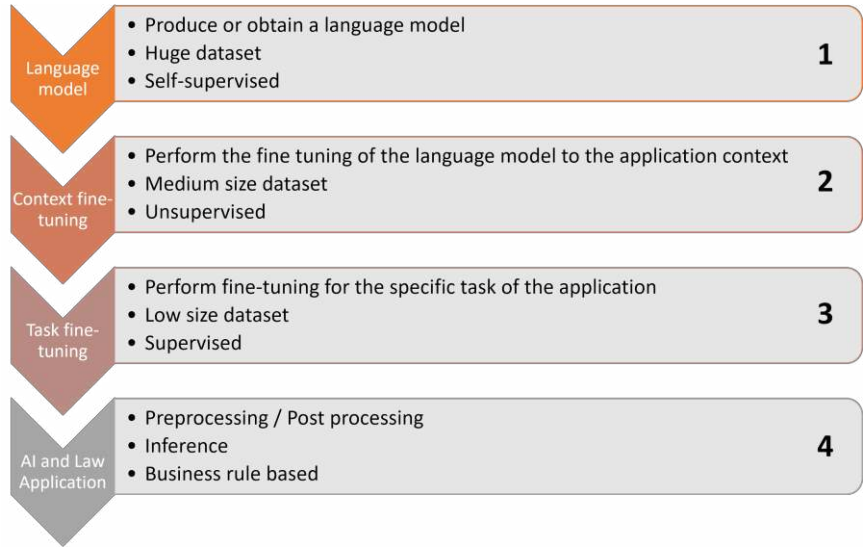

Figure 2. Technical aspects of the proposed methodology

2.1. Language Model

The first layer involves producing or obtaining a pre-trained language model (PLM) able to process texts in Portuguese, also known as foundation models (Bai *et al.,* 2023; Bommasani *et al.,* 2022). This step is entirely dependent on the language, since PLMs are pre-trained in language specific texts, thus the availability and quality of these models is language dependent. Some languages like English and German have much more and bigger language models already available. At this moment, in Portuguese there are models in BERT (Devlin *et al.,* 2019), GPT-2 (Radford *et al.,* 2019) and T5 (Raffel *et al.,* 2019) transformer-based architectures. More recently, with the emergence of Large Language Models (LLMs), several models are being trained with texts from different languages, called multilingual models (Qin *et al.,* 2024), which also can be used in this layer of the proposed methodology. Examples of multilingual LLMs are mT5 (Xue *et al.,* 2021), Mixtral (Jiang *et al.,* 2024), Llama (Touvron *et al.,* 2023), GPT-4 (Openai *et al.,* 2024), among very others.

Training a language model from scratch is a resource-intensive process, involving large datasets and substantial computational power, typically using Graphics Processing Unit (GPU) or Tensor Processing Unit (TPU). This makes the training process both expensive and environmentally taxing. The general trend to achieve improved performance in language models is to increase their size and the volume of training data. This need for large datasets and model architecture may prompt the decision to train a new model, particularly if the chosen architecture lacks a pre-existing Portuguese model.

In most cases, however, the production of a new language model in Portuguese is not required, unless a particular model architecture is necessary but lacks available pre-trained models in the language. Existing Portuguese models such as BERT, T5, and GPT-2, as well as the multilingual LLMs provide strong alternatives for most applications.

RECIMA21 - REVISTA CIENTÍFICA MULTIDISCIPLINAR
ISSN 2675-6218

Furthermore, an important consideration when training a model from scratch would be to utilize more extensive and domain-specific datasets. For example, BERTimbau (Souza; Nogueira; Lotufo, 2020), the first pre-trained BERT model for Brazilian Portuguese, was trained on the Brazilian Portuguese brWaC Corpus (Wagner Filho *et al.,* 2018), which includes 145 million sentences and 2.7 billion tokens from .br domain pages.

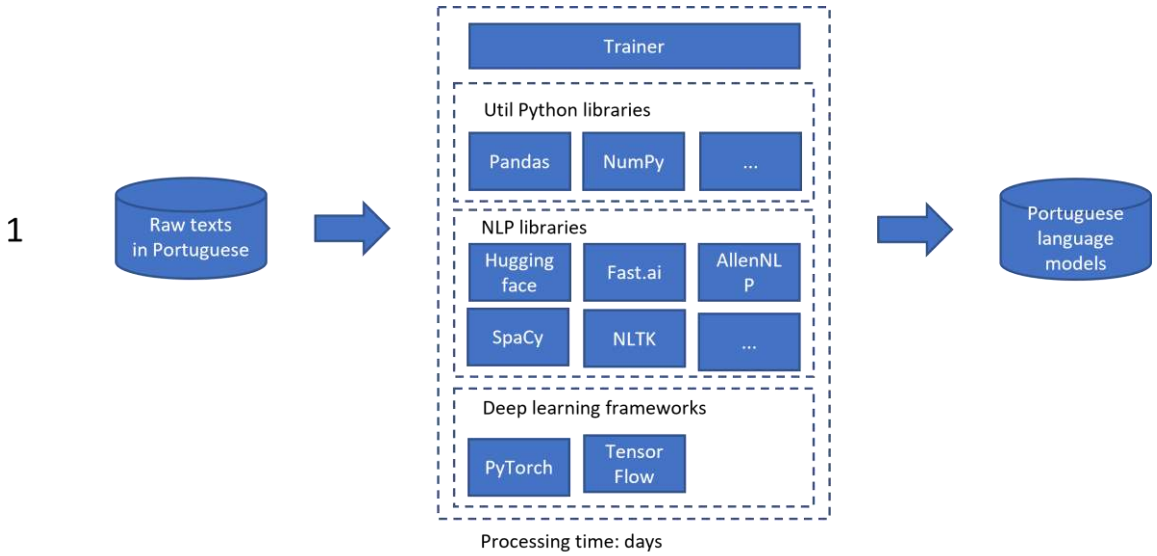

Figure 3. Training a language model from scratch

Figure 3 illustrates the components needed to train a language model from scratch. In this case, the input is raw texts in Portuguese and the output is a Portuguese language model. The processing is performed by a trainer component, usually developed using a Python technology stack, including from the basis to the top: a deep learning framework, usually PyTorch or TensorFlow; NLP libraries as HuggingFace, Fast.ai, AllenNLP, SpaCy, etc., and others util Python libraries like Pandas, NumPy, etc. The figure is not comprehensive and does not include all the software components needed and technical details for training a language model, like the tokenizer, which is a very important component to do this job, for example.

Although it is possible to train a language model from scratch using domain-specific data (such as legal texts), this approach is uncommon. More typically, fine-tuning an existing model on a specific context, as discussed in section Context fine-tuning, is preferred.

RECIMA21 - REVISTA CIENTÍFICA MULTIDISCIPLINAR
ISSN 2675-6218

2.2. Context Fine-Tuning

The second layer involves performing the fine-tuning of the pre-trained language model in the application context. The objective is to improve a general Portuguese language model with the nuances of the application context where it will be used. For example, Law field applications should be context fine-tuning with texts from court proceedings. More specifically, if the application where the final model will be used is going to process lawyer texts, the language model should be fine-tuned with petition texts. On the other hand, if the application will process texts written by judges, the language model should be fine-tuned with texts from orders and sentences. Or both, if the application will process texts from lawyers and judges.

In this layer, the language model undergoes additional unsupervised training, typically using a moderate amount of raw text. Although this process requires less computational power than the initial large-scale pre-training, it still demands significant resources, including machines equipped with GPUs or TPUs. While optional, context-specific fine-tuning can significantly enhance the model's performance on downstream NLP tasks. Moreover, this fine-tuning approach is highly versatile, extending beyond the legal domain and demonstrating value across a wide range of NLP applications.

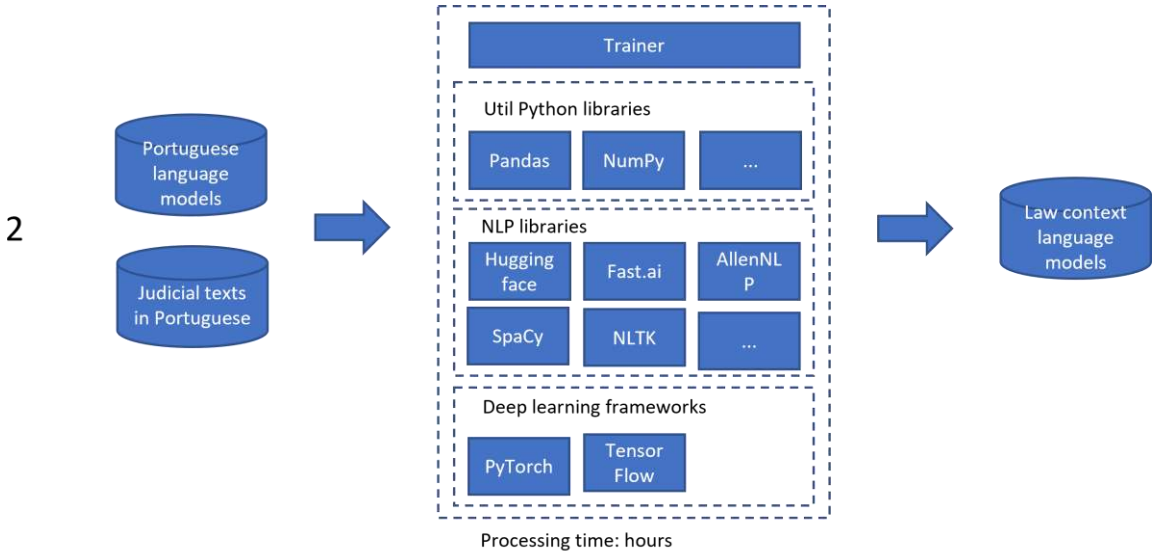

Figure 4. Context fine-tuning

Figure 4 illustrates the context fine-tuning process, where raw judicial texts in Portuguese are used alongside the general Portuguese language model from the first step. The output is a domain-specific language model, adapted to the legal context. The fine-tuning process is carried out by a trainer component similar to the one used in the first layer.

RECIMA21 - REVISTA CIENTÍFICA MULTIDISCIPLINAR
ISSN 2675-6218

2.3. Task Fine-Tuning

The third layer of the proposed methodology focuses on fine-tuning a PLM for a specific NLP downstream task, tailored to the application's requirements. These tasks might include sequence classification, named entity recognition (NER), summarization, or other specialized activities. The input to this layer is the output from one of the earlier stages: either a general Portuguese or multilingual language model (first layer) or a legal context-specific language model (second layer).

At this stage, supervised learning is employed using annotated datasets, which are often relatively small in size. This step is considered essential for non-large language models (non-LLMs), such as BERT or T5, given the highly specific nature of the legal field. Legal applications demand specialized models with high accuracy to be effective in real-world scenarios. However, for LLMs, task-specific fine-tuning can sometimes be bypassed due to their inherent ability to generalize across a wide range of tasks using zero-shot learning (Pushp; Srivastava, 2017). When skipping task-specific fine-tuning for LLMs, subsequent stages will require meticulous prompt engineering to guide the model effectively. Techniques such as Retrieval-Augmented Generation (RAG) (Gao *et al.,* 2024) and few-shot learning (Wang *et al.,* 2020) can help compensate for the absence of task-level training, which remains critical for smaller models.

Annotated datasets are central to this stage. For example, the LeNER_Br dataset (Luz de Araujo *et al.,* 2018) can be used for NER tasks involving judicial texts.

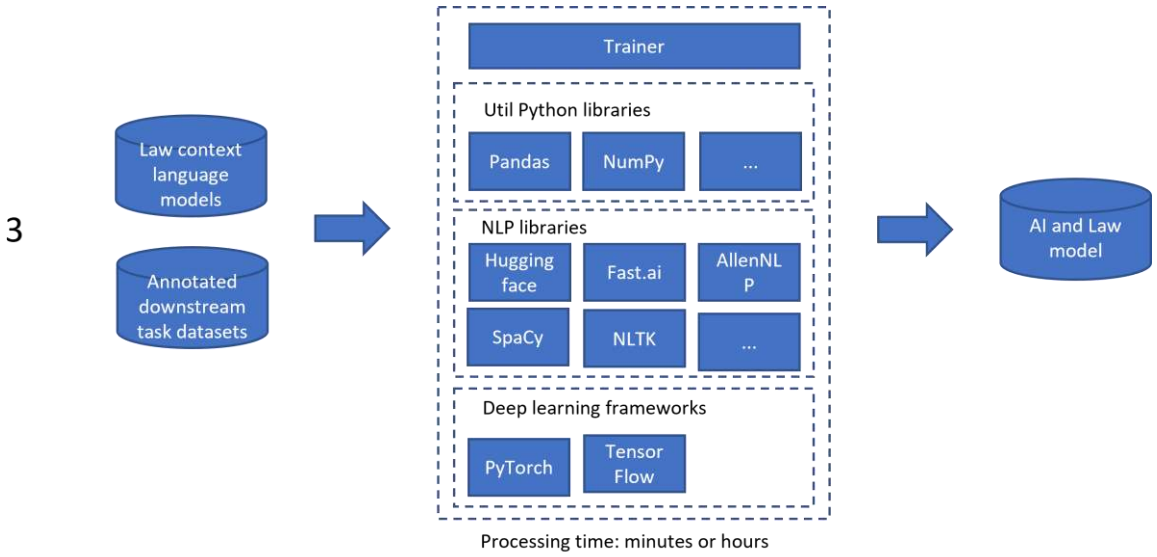

Figure 5. Task fine-tuning

Figure 5 illustrates the task fine-tuning process. Unlike earlier layers, the input here is an annotated dataset specific to the downstream task, rather than raw text. The model to be fine-tuned is either the general Portuguese language model from the first step or the judicial context-specific language model from the second step, where applicable. The output is an AI-and-Law model: a task-

RECIMA21 - REVISTA CIENTÍFICA MULTIDISCIPLINAR
ISSN 2675-6218

specific NLP model designed to handle judicial tasks in Portuguese. The fine-tuning process leverages the same training infrastructure used in the second step, ensuring consistency and efficiency.

This layer represents a pivotal step in adapting PLMs to the unique demands of legal applications, ensuring both accuracy and relevance in specialized tasks while building reusable, high-quality models for broader application.

2.4. AI and Law Application

The final layer of the proposed methodology focuses on deploying the fine-tuned AI model from the previous layer in real-world applications, enabling inference and practical use. In this context, the AI model generates predictions based on input data, and these predictions are integrated with additional processing to add value within the specific business context. For instance, in a NER task, the fine-tuned model identifies entities in legal texts, and the application processes these results to support decision-making or automate tasks.

The fourth and final layer of the methodology, illustrated in ***Figure 6***, focuses on the deployment of the AI model for inference and practical application. Part (a) of the figure outlines the key components required for this process. The inputs to this stage are the AI and Law model fine-tuned in the third layer and the application data, typically text inputs provided by users or systems. The output, referred to as the AI and Law response, represents the AI-generated results tailored for user consumption or further system integration.

This layer encompasses three core elements: preprocessing, AI inference, and postprocessing. Preprocessing ensures the input data aligns with the model's requirements by applying business rules to clean, format, or segment the text. Postprocessing, on the other hand, adapts the AI output to align with business needs, enhancing its usability and relevance for the specific context.

The AI inference component serves as the core processing unit, performing predictions using the fine-tuned model. This component may closely resemble the trainer module used in earlier steps but is optimized for inference tasks. It can be directly embedded within the application's architecture or deployed as a standalone API, offering flexibility for integration and scalability.

Part (b) of ***Figure 6*** illustrates how AI results are delivered to users, emphasizing their integration into accessible interfaces or automated pipelines.

RECIMA21 - REVISTA CIENTÍFICA MULTIDISCIPLINAR
ISSN 2675-6218

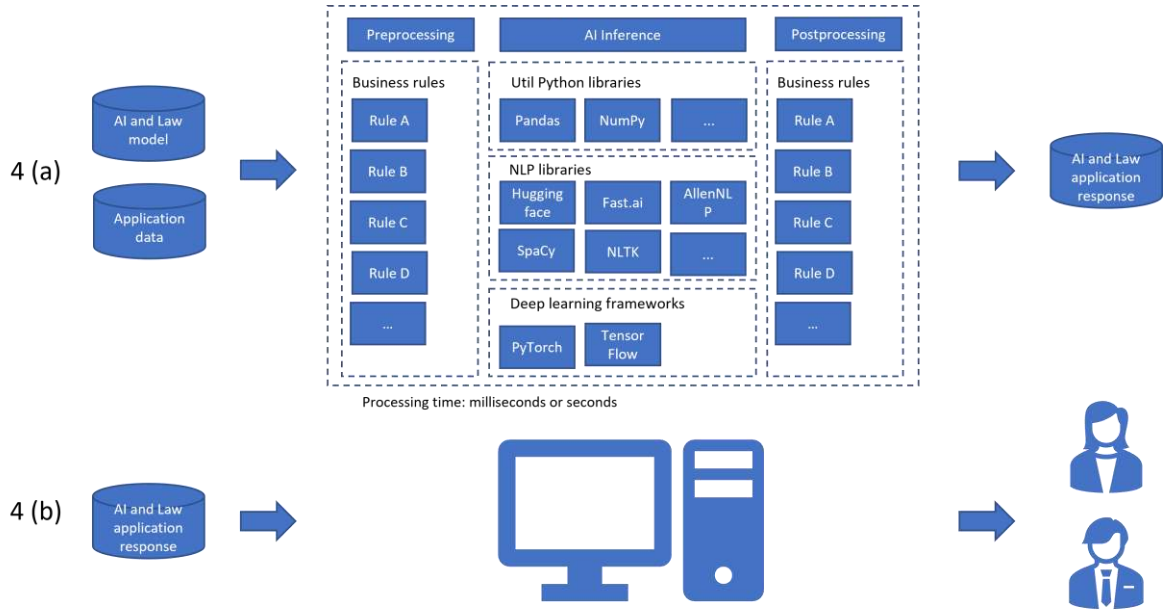

Figure 6. AI and Law application

While the training phase requires substantial computational resources, typically leveraging GPUs or TPUs to handle large datasets and complex calculations, the inference phase is less computationally intensive. However, GPUs can still enhance performance during inference by processing large amounts of data concurrently. For environments constrained to CPU usage, optimization techniques for transformer-based models are being actively developed to improve inference efficiency (Dice; Kogan, 2021; Hsu *et al.,* 2020).

For LLMs, inference differs slightly due to their ability to handle significantly larger inputs, known as the context window. Modern LLMs support context lengths of up to 128k tokens or more, enabling applications such as summarizing entire books or processing extensive legal documents in a single inference call[6][7]. Additionally, innovations such as infinite context windows are emerging, further expanding the capabilities of these models (Munkhdalai; Faruqui; Gopal, 2024).

This phase goes beyond mere AI inference and emphasizes aligning AI outputs with business rules and user needs. To ensure the solution effectively addresses business requirements, preprocessing and postprocessing are critical steps. Preprocessing prepares the input text by removing unnecessary components, such as headers, footers, or irrelevant sections. Postprocessing adapts the AI-generated response for practical use, improving clarity, usability, and accuracy within the specific business scenario.

[6] https://platform.openai.com/docs/models
[7] https://artificialanalysis.ai/leaderboards/models

RECIMA21 - REVISTA CIENTÍFICA MULTIDISCIPLINAR
ISSN 2675-6218

While transformer-based models like BERT have reduced the dependency on extensive NLP preprocessing (Özçift *et al.,* 2021), it remains essential in practical applications, particularly in business contexts. For example, preprocessing might involve:

- Removing HTML tags, special characters, or unnecessary whitespaces.
- Converting text to lowercase (if appropriate), expanding contractions, or normalizing numbers.
- Segmenting lengthy documents into smaller chunks to fit within the model's token limit.

However, preprocessing must be tailored to the application's domain. For legal texts, excessive preprocessing, such as lowercasing or removing stopwords, may inadvertently alter the meaning or context of the input.

Postprocessing ensures that AI outputs are ready for consumption by either end-users or other software systems. For example, NER model responses often require further refinement. ***Figure 7***and ***Figure 8*** illustrate the difference between raw NER output and postprocessed results, highlighting the importance of enhancing usability for downstream applications.



Figure 7. NER response with postprocessing

RECIMA21 - REVISTA CIENTÍFICA MULTIDISCIPLINAR
ISSN 2675-6218

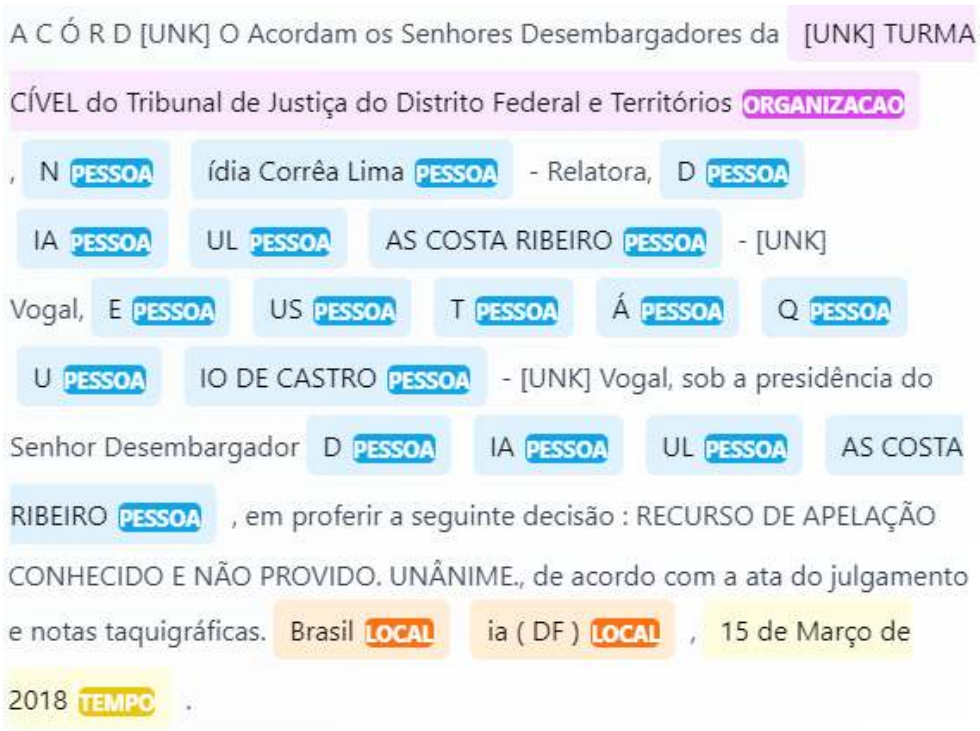

Figure 8. NER response without postprocessing

2.5. Example Set of Components

In addition to outlining the steps of the proposed methodology for building AI and Law applications, this work also presents a view of an example set of components for the methodology, illustrated in ***Figure 9***. These components are divided into layers: data, model, context adaptation, task, application, and extensions. The figure also shows the methodology layers (1 to 4) for clarity. New components will be developed in the future to expand the methodology and support AI and Law application development.

The foundation of the stack is the data layer, which houses Portuguese corpora used as the basis for building PLMs. These corpora primarily consist of raw Portuguese text, often collected from various internet sources. Among the most notable resources is the brWaC Corpus, which includes 145 million sentences, and 2.7 billion tokens derived from 38 million .br domain pages. This corpus underpins Portuguese PLMs such as BERTimbau (Souza; Nogueira; Lotufo, 2020) and PTT5 (Carmo *et al.,* 2020), adaptations of BERT (Devlin *et al.,* 2019) and T5 (Raffel *et al.,* 2019), respectively. Another important resource is the OSCAR corpus, a massive multilingual dataset extracted from the Common Crawl, containing 10.7 billion words. This corpus represents a rich resource for training new PLMs. Data from this layer serves as input to the first step of the proposed methodology.

RECIMA21 - REVISTA CIENTÍFICA MULTIDISCIPLINAR
ISSN 2675-6218



Figure 9. Stack of the methodology components

The model layer encompasses PLMs in Portuguese, built on transformer-based architectures. Key examples include BERTimbau, PTT5, and GPT-2-small[8]. However, there is still a need for more diverse PLMs, including models trained on alternative corpora such as OSCAR. This layer provides outputs for the first step of the methodology and forms part of the input for the second step.

The context adaptation layer focuses on adapting PLMs to the judicial domain through fine-tuning on specialized datasets, enriching the model with domain-specific knowledge. Judicial texts, which are often underrepresented in general corpora like Wikipedia or Common Crawl, are critical in this context.

This layer includes both judicial datasets and adapted models. Examples of datasets include the Acordaos TCU corpus[9], the Iudicium Textum Dataset (Willian Sousa; Fabro, 2019) and a private lawyer's petition dataset. Examples of adapted models include those listed in Results section.

Data from this layer, along with models from the previous layer, serve as input for the second step, while the adapted models become inputs for the third step.

The task layer includes resources for specific NLP downstream tasks, such as named entity recognition (NER) or summarization. These resources consist of annotated datasets for fine-tuning,

[8] https://huggingface.co/pierreguillou/gpt2-small-portuguese
[9] https://www.kaggle.com/ferraz/acordaos-tcu

RECIMA21 - REVISTA CIENTÍFICA MULTIDISCIPLINAR
ISSN 2675-6218

such as LeNER_Br (Luz de Araujo *et al.,* 2018), and task-specific fine-tuned models that are ready for use.

Whenever possible, leveraging PLMs from this layer is encouraged to maximize reuse and minimize the need for retraining. Data from this layer, combined with context-adapted models, serves as input to the third step, while task-specific models are outputs of this step and inputs to the fourth step.

The application layer incorporates business rules for both preprocessing input texts and postprocessing model outputs. While specific to each application, some common patterns can be generalized to ensure usability and effectiveness in real-world scenarios.

Inputs to this layer include application data and task-specific models, along with relevant business rules. The output is the AI and Law application response, the final product consumed by end-users or other systems.

The extensions layer introduces optional, but highly recommended, components for enhancing AI and Law applications. These include:

- Explainable AI. Essential for fostering trust and accountability in legal applications (Adadi; Berrada, 2018).
- Judicial Ontologies. Formal descriptions of legal concepts that can enrich language models (Noy; Mcguinness, 2001).
- AI and Law Application Metric. A novel evaluation metric tailored to assess the application-level success of AI in achieving real-world objectives.

Future developments will likely expand this layer with additional components, further enhancing the methodology.

By combining these layers, the proposed stack provides a structured and flexible framework for developing robust AI and Law applications, ensuring adaptability to evolving business and technological needs.

3. RESULTS

The proposed methodology was successfully applied in real-world scenarios at the Court of Justice of Rio Grande do Sul, demonstrating its practical value. A key application was the Judgment Report Generator, which automated the creation of judgment reports using Generative AI, a groundbreaking innovation in judicial workflows. Evaluation employed a unique multi-layered approach, combining traditional NLP metrics, human assessments by judges, and contextual evaluations with application-specific datasets. Notably, a fine-tuned GPT-4o-Mini model outperformed GPT-4o in this legal domain, emphasizing the importance of task-specific validation over generic benchmarks.

Other experiments involved fine-tuning BERT and GPT-2-based Portuguese language models for tasks like legal Named Entity Recognition (NER) and context-adapted text generation. Fine-tuned models achieved state-of-the-art performance on the LeNER-Br dataset (Zanuz; Rigo, 2022) and

RECIMA21 - REVISTA CIENTÍFICA MULTIDISCIPLINAR
ISSN 2675-6218

generated context-specific responses tailored to judicial or legal texts. These experiments validated the effectiveness of transfer learning with minimal data, highlighting the adaptability of pre-trained models to diverse legal contexts.

The results underscore the methodology's potential to align cutting-edge NLP techniques with judicial needs, offering reliable, scalable, and context-aware AI solutions. By addressing challenges like dataset specificity, explainability, and regulatory compliance, the study establishes a robust foundation for advancing AI in the legal domain.

4. CONCLUSION AND FUTURE WORK

This paper presented a comprehensive methodology for developing AI applications in the legal domain, leveraging state-of-the-art NLP technologies such as transformer-based architectures, pre-trained language models, transfer learning, and LLMs. The framework integrates recent advances in NLP to address the unique needs of the judicial field, enabling the creation of reliable, scalable, and high-quality applications while simplifying development for practitioners.

The methodology highlights the importance of collaboration between Business and IT Teams in defining success metrics, creating datasets, and refining prototypes like POCs and MVPs. This iterative process ensures alignment with business objectives and the accuracy required in legal contexts. Key components such as explainable AI, pre-trained models, and application-level metrics further enhance its technical rigor and practical relevance.

The methodology aligns with Brazilian Resolution No. 332 of the National Council of Justice (CNJ), which mandates transparency, explainability, and governance in judicial AI systems. Anticipating updates for LLMs and generative AI ensures compliance with evolving regulations, supporting practical application and ethical adoption in real-world scenarios.

Future work will refine optional components like judicial ontologies, improved explainability features, and a new application-level evaluation metric to enhance transparency, adaptability, and effectiveness. These innovations aim to establish a strong standard for AI systems in the legal domain, bridging NLP advancements with judicial needs.

REFERENCES

ADADI, A.; BERRADA, M. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). **IEEE Access**, v. 6, p. 52138–52160, 2018.

BAI, Y. *et al*. Benchmarking Foundation Models with Language-Model-as-an-Examiner. **Advances in Neural Information Processing Systems**, v. 36, p. 78142–78167, 15 dez. 2023.

BOMMASANI, R. *et al*. On the Opportunities and Risks of Foundation Models. **arXiv**, 12 jul. 2022. Available at: https://doi.org/10.48550/arXiv.2108.07258. Available at: 26 out. 2024

CARMO, D. *et al.* PTT5: Pretraining and validating the T5 model on Brazilian Portuguese data. **arXiv**, 8 out. 2020.

RECIMA21 - REVISTA CIENTÍFICA MULTIDISCIPLINAR
ISSN 2675-6218

CLARK, P. *et al.* Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge. **arXiv**, 14 mar. 2018. Available at: https://www.doi.org/10.48550/arXiv.1803.05457. Available at: 24 out. 2024

DEVLIN, J. *et al.* BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. **Proceedings of the 2019** […] Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). NAACL-HLT 2019. Minneapolis, Minnesota: Association for Computational Linguistics, jun. 2019. Available at: http://doi.org/10.18653/v1/N19-1423. Available at: 6 ago. 2021

DICE, D.; KOGAN, A. Optimizing Inference Performance of Transformers on CPUs. **arXiv**, 22 fev. 2021.

GAO, Y. *et al.* Retrieval-Augmented Generation for Large Language Models: A Survey. **arXiv**, 27 mar. 2024. Available at: https://doi.org/10.48550/arXiv.2312.10997. Available at: 27 out. 2024

HENDRYCKS, D. *et al.* Measuring Massive Multitask Language Understanding. **arXiv**, 12 jan. 2021. Available at: http://doi.org/10.48550/arXiv.2009.03300. Available at: 24 out. 2024

HSU, Y.-T. *et al*. Efficient Inference For Neural Machine Translation. **Proceedings of SustaiNLP**: Workshop on Simple and Efficient Natural Language Processing EMNLP-SUSTAINLP 2020. Online: Association for Computational Linguistics, nov. 2020. Available at: https://www.doi.org/10.18653/v1/2020.sustainlp-1.7. Available at: 20 fev. 2022

JIANG, A. Q. *et al*. Mixtral of Experts. **arXiv**, 8 jan. 2024. Available at: https://doi.org/10.48550/arXiv.2401.04088. Available at: 26 out. 2024

LIN, C.-Y. ROUGE: A Package for Automatic Evaluation of Summaries. Text Summarization Branches Out. **Anais** [...] Barcelona, Spain: Association for Computational Linguistics, jul. 2004. Available at: https://aclanthology.org/W04-1013. Available at: 25 jun. 2024

LUZ DE ARAUJO, P. H. *et al.* LeNER-Br: A Dataset for Named Entity Recognition in Brazilian Legal Text. *In*: VILLAVICENCIO, A. et al. (Eds.). **Computational Processing of the Portuguese Language**. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2018. v. 11122. p. 313–323.

MUNKHDALAI, T.; FARUQUI, M.; GOPAL, S. Leave No Context Behind: Efficient Infinite Context Transformers with Infini-attention. **arXiv**, 9 ago. 2024. Available at: https://doi.org/10.48550/arXiv.2404.07143. Available at: 27 out. 2024

NOY, N. F.; MCGUINNESS, D. L. **Ontology Development 101:** A Guide to Creating Your First Ontology. [S. l.] Stanford Knowledge Systems Laboratory, 2001.

OPENAI et al. GPT-4 Technical Report. **arXiv**, 4 mar. 2024. Available at: https://doi.org/10.48550/arXiv.2303.08774. Available at: 4 ago. 2024

ÖZÇIFT, A. *et al.* Advancing natural language processing (NLP) applications of morphologically rich languages with bidirectional encoder representations from transformers (BERT): an empirical case study for Turkish. **Automatika**, v. 62, n. 2, p. 226–238, 3 abr. 2021.

PAN, S. J.; YANG, Q. A Survey on Transfer Learning. **IEEE Transactions on Knowledge and Data Engineering**, v. 22, n. 10, p. 1345–1359, out. 2010.

PAPINENI, K. *et al.* BLEU: a method for automatic evaluation of machine translation. **Proceedings of the 40th Annual Meeting on Association for Computational Linguistics**. ACL' 02.USA:

RECIMA21 - REVISTA CIENTÍFICA MULTIDISCIPLINAR
ISSN 2675-6218

Association for Computational Linguistics, 6 jul. 2002. Available at: https://dl.acm.org/doi/10.3115/1073083.1073135. Access at: 4 ago. 2024

PUSHP, P. K.; SRIVASTAVA, M. M. Train Once, Test Anywhere: Zero-Shot Learning for Text Classification. **arXiv**, 23 dez. 2017.

QIN, L. *et al*. Multilingual Large Language Model: A Survey of Resources, Taxonomy and Frontiers. **arXiv**, 7 abr. 2024. Available at: https://doi.org/10.48550/arXiv.2404.04925. Access at: 26 out. 2024

RADFORD, A. *et al*. Improving language understanding by generative pre-training. **OpenAI blog**, 2018.

RADFORD, A. *et al*. Language models are unsupervised multitask learners. **OpenAI blog**, v. 1, n. 8, p. 9, 2019.

RAFFEL, C. *et al*. Exploring the limits of transfer learning with a unified text-to-text transformer. **arXiv preprint arXiv:1910.10683**, 2019.

SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. BERTimbau: Pretrained BERT Models for Brazilian Portuguese. Intelligent Systems. **Anais** [...] Lecture Notes in Computer Science.Cham: Springer International Publishing, 2020. Available at: http://doi.org/10.1007/978-3-030-61377-8_28

TOUVRON, H. *et al.* LLaMA: Open and Efficient Foundation Language Models. **arXiv**, 27 fev. 2023. Available at: https://doi.org/10.48550/arXiv.2302.13971. Access at: 29 mar. 2024

VASWANI, A. *et al.* Attention is all you need. **Proceedings of the 31st International Conference on Neural Information Processing Systems**. NIPS'17.Red Hook, NY, USA: Curran Associates Inc., 4 dez. 2017. Available at: https://dl.acm.org/doi/10.5555/3295222.3295349. Access at: 6 ago. 2021

WAGNER FILHO, J. A. *et al.* The brWaC Corpus: A New Open Resource for Brazilian Portuguese. **Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).** Miyazaki, Japan: European Language Resources Association (ELRA), maio 2018. Available at: https://aclanthology.org/L18-1686. Access at: 17 ago. 2021

WANG, Y. *et al.* Generalizing from a Few Examples: A Survey on Few-shot Learning. **ACM Computing Surveys**, v. 53, n. 3, p. 63:1-63:34, 12 jun. 2020.

WILLIAN SOUSA, A.; FABRO, M. Iudicium Textum Dataset Uma Base de Textos Jurídicos para NLP. 2nd Dataset Showcase Workshop at SBBD (Brazilian Symposium on Databases). **Anais** [...] Em: SBBD 2019. Fortaleza, Brazil: 7 out. 2019.

XUE, L. *et al*. mT5: A massively multilingual pre-trained text-to-text transformer. **arXiv**, 11 mar. 2021. Available at: https://doi.org/10.48550/arXiv.2010.11934. Access at: 26 out. 2024

ZANUZ, L.; RIGO, S. J. Fostering Judiciary Applications with New Fine-Tuned Models for Legal Named Entity Recognition in Portuguese. (V. Pinheiro et al., Eds.)Computational **Processing of the Portuguese Language**. Cham: Springer International Publishing, 2022. Available at: https://doi.org/10.1007/978-3-030-98305-5_21

ZELLERS, R. et al. HellaSwag: Can a Machine Really Finish Your Sentence?. **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**. ACL 2019. Florence, Italy: Association for Computational Linguistics, jul. 2019. Available at: http://doi.org/10.18653/v1/P19-1472. Access at: 24 out. 2024

ZHENG, L. *et al.* Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. **Advances in Neural Information Processing Systems**, v. 36, p. 46595–46623, 15 dez. 2023.