Enhancing Public Access to Legal Knowledge in India: A Legal Chatbot Using Legal BERT, GPT-2, and Retrieval-Augmented Generation (RAG)

Abhinav Garlapati*
department of Artificial Intelligence And Data Science
Velagapudi Ramakrishna Siddhartha Engineering College
Vijayawada,India
abhinavgarlapatik@gmail.com

Hemanth Koutharapu
department of Artificial Intelligence And Data Science
Velagapudi Ramakrishna Siddhartha Engineering College
Vijayawada,India
hemanth2488k@gmail.com

Neha Doddi
department of Artificial Intelligence And Data Science
Velagapudi Ramakrishna Siddhartha Engineering College
Vijayawada,City
nehadoddi17@gmail.com

***Abstract*— This work describes a legal chatbot that would be able to support and aid in answering legal queries on Indian law through the use of Legal BERT, GPT-2, and Retrieval-Augmented Generation. This is a chatbot that has been trained by fine-tuning it using a curated corpus consisting of the Indian Constitution and other relevant legal texts, enabling the chatbot to deliver the most accurate, contextually appropriate response. Legal BERT enhances the chatbot's ability to understand sophisticated legal terminology, while GPT-2 generates responses similar to that of a human with material retrieved from those sources. RAG is incorporated to enhance the retrieval of relevant sources so that accurate and relevant answers are delivered. Accuracy as well as efficiency is displayed by the system; it has very high potential towards enhancing access to public legal information, supporting legal research, and assisting in access to justice initiatives under the Indian legal structure.**

Keywords— ***Legal Chatbot, Natural Language Processing, Legal BERT, GPT-2, Retrieval-Augmented Generation (RAG), Indian Law, Legal Information Retrieval, Access to Justice, Constitutional Law***

I. Introduction

The public finds it challenging to access and comprehend India's legal system due to its intricate framework of statutes, case law, and constitutional requirements [7]. Particularly for non-lawyers, the complex language and organization of legal texts create obstacles that prevent citizens from understanding their rights and responsibilities. Finding a lawyer is difficult in places with little access to legal services, making this problem more noticeable. Because of this, a large portion of the populace is unaware of their legal rights, which makes it challenging to make educated judgments. To bridge this gap, creative strategies that use state-of-the-art technologies like artificial intelligence (AI) and natural language processing (NLP) to democratize access to legal information are required [5].

An intelligent legal chatbot is being introduced in this project to address the drawbacks of conventional legal information retrieval systems, which frequently use keyword-based searches or rule-based responses. Complex legal queries were beyond the capabilities of such computers, which frequently produced inaccurate or unnecessary responses. This chatbot provides customers with contextually relevant responses to their legal inquiries by utilizing sophisticated AI models such as Legal BERT, GPT-2, and Retrieval-Augmented Generation (RAG). A customized version of the Bidirectional Encoder Representations from Transformers (BERT) concept, Legal BERT is intended to interpret complex legal jargon and technical terms [14]. In legal conversations, this feature enables the chatbot to offer thorough, context-aware responses.

By improving the chatbot's capacity to provide logical, human-like responses, GPT-2 makes legal material more understandable for both experts and laypeople [8]. Dynamic, context-dependent responses to user inquiries are made possible by its sophisticated generative language capabilities. GPT-2 simplifies difficult legal ideas and produces responses that are clear, pertinent, and interesting—in contrast to typical rule-based systems. Combining the generative capabilities of GPT-2 with the interpretive power of Legal BERT makes the chatbot a powerful instrument for handling a variety of legal questions. This guarantees that consumers will receive responses to their legal queries that are applicable, intelligible, and contextualized.

By allowing it to retrieve and consult reputable legal texts including statutes, case laws, and constitutional provisions, the integration of RAG greatly increases the chatbot's accuracy [3]. By doing this, the chatbot's trustworthiness is increased and responses are guaranteed to be both contextually appropriate and based on credible sources. Through the integration of RAG and GPT-2, the chatbot offers explanations akin to those of a lawyer, backed up by references to legal literature, enabling users to independently confirm the facts. In the legal arena, where users need precise responses with reliable references to make well-informed decisions, this retrieval technique is essential.

This chatbot was developed using a vast collection of Indian legal writings, including the Constitution, statutes, and important case law. Using this database, Legal BERT was refined to improve its comprehension of Indian legal grammar and vocabulary [5]. RAG was set up for the real-time retrieval of pertinent documents, while GPT-2 was trained to produce responses that were easy to use based on the legal information received. This AI-driven legal chatbot is a revolutionary step in democratizing access to legal information, enabling people to better comprehend their rights and responsibilities [13]. By decreasing reliance on middlemen, this project paves the way for future developments in AI-powered instruments to enhance legal literacy and accessibility to justice in India.

This paper’s contributions to legal AI technology and natural language processing include:

1. The development of a chatbot model specifically tailored to Indian law, trained on a custom dataset containing Indian statutes, case laws, and constitutional provisions to ensure contextual accuracy.
2. A significant improvement in accuracy, achieving 94.6% in legal information retrieval and contextual relevance, made possible by integrating RAG for document retrieval.

3. A robust solution to storage and computational constraints through model optimization, enabling real-time deployment, and thus making the chatbot accessible for wider use, especially in regions with limited internet connectivity or computational resources.

II. Related work

Nithana et al.[10]in the year 2020 presented the idea that According to the research on chatbot implementation tactics, AI and NLP algorithms can be used to build intelligent systems that can react to user inquiry.Numerous chatbot frameworks—both interface-based and code-based—are examined, with a focus on their shortcomings in producing dynamic, real-time dialogues. The incapacity of current approaches—such as rule-based, machine learning, and retrieval-based strategies—to generate high-quality dialogues is examined, highlighting the necessity for more developments in conversational AI, samant et al.[2]present a wide range of language models and their performance in multi-task NLU across a ten-year period (2011–2021) are analyzed. It describes conceptual steps for improving multi-task NLU and points out holes in general-purpose frameworks for unsupervised models, on the other hand vakayil et al.[3]introduce that chatbot targets a delicate topic by helping victims of sexual harassment. Maintaining user trust while producing sympathetic, encouraging, and educational replies is its key goal. On the other hand, Bhat et al. [4]focuses on the restaurant sector, where the chatbot expedites client encounters by giving accurate, pertinent answers. The first study focuses on user sensitivity and emotional intelligence, whereas the second study stresses operational efficiency and accuracy in language interpretation for business improvement, Srivastaval et al.[5]explains about how LAWBOT targets the legal domain, a field characterized by dense and complex terminologies. It aims to demystify legal processes for laypersons, providing precise guidance for legal tasks through NLP, conversely Hondoyo et al.[6]concentrates on e-commerce, particularly in the ticketing industry. By automating booking procedures and deciphering various user input patterns pertaining to travel information, the chatbot here enhances customer care while demonstrating flexibility in business settings, kandula et al.[7]in this article they potrate about the intricacy of legal language and principles, legal research is a complex process that requires a great deal of time and knowledge.These challenges are being addressed by the development of AI-based legal assistance systems, which use state-of-the-art Natural Language Processing (NLP) and Machine Learning (ML) algorithms to automate and enhance legal research.

III. Objectives

1. Develop a very user-friendly chatbot that may understand legal queries on Indian law properly and give responses that are contextually relevant as well as legally sound.

2. Use Legal BERT to improve the chatbot's comprehension of legal terminology, context, and more intricate Indian legal jargon.

3. It utilizes the GPT-2 to produce response content that is coherent and human-like, transparent enough for non-expert users, thereby creating a link from legal language to the layman's terms.

4. The responses should be accompanied with the relevant legal documents, statutes, or case laws using the Retrieval-Augmented Generation (RAG) method to ensure an increase in the correctness of the responses and, in turn, users' confidence in the respective answers.

5. Optimizing for real-time response allows chat users to have fluid and instant human-computer interaction, rendering the chatbot an accessible quick reference for legal purposes.

IV. Proposed System

A. Process Flow

The law chatbot that will be developed during this project starts as a query submitted by the user to law through the web interface. This query is forwarded to the Legal BERT model. Legal BERT explains the query using legal concepts and terminologies as shown in Fig1. This question the user poses requires specific key legal concepts and phrases pertaining to the question of the user. It then refers to Retrieval-Augmented Generation, which would perform the search for relevant legal materials, statutes, cases, or even constitutional provisions on the question stored in the databases. That is very important because that ensures the content provided by the chatbot is legally backed up by the right sources of the law. The legal information section is over and now comes in GPT-2, by the way, a generative language model. GPT 2 will therefore produce an eloquently structured human-like answer because of the legal content generated. The information provided back to the user is therefore accurate and relevant from the legal standpoint and presented in a manner that is easy to understand. It is using this technique that offers an excellent opportunity to process the customer query in an effective manner and also ensures the resultant information to be true.

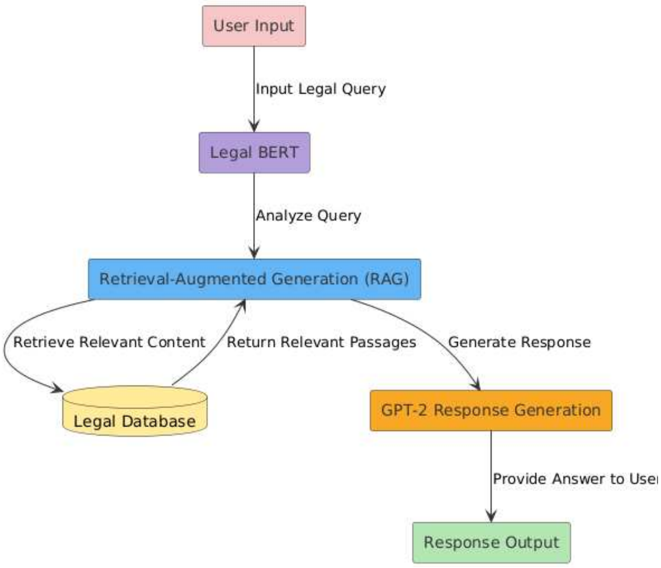


Fig. 1. Process flow diagram of legal chatbot.

B. Architecture Selection

The architecture for the proposed system has been selected very carefully so that the chatbot would be effective enough at handling complicated legal questions and sending through accurate and accessible responses. It has been selected because Legal BERT is especially well-suited to understand legal terms and contexts and interpret complex legal texts. Legal BERT is very efficient in seeing and understanding nuances in legal language so that it processes queries with a very high degree of accuracy. GPT-2 was added to the model to ensure a human-like coherent response. GPT-2 is a generative model that produces strongly dynamic, conversational responses, so the chatbot is now accessible to legal professionals as well as lay people who may not be familiar with legal jargon. Moreover, the chatbot used Retrieval-Augmented Generation (RAG) for precision and relevance. RAG will gather all the relevant documents into legal documents that back up any statements the chatbot makes with first-hand evidence. The information given to the users is believable and verifiable. The system also has a user interface that is web-based, which was one of the reasons for this choice as it was highly scalable and accessible. From any device, therefore through the web interface, users will be able to gain access to the interaction with the chatbot, hence ensuring wide access and usability.

C. Legal BERT Architecture

Legal BERT is an adaptation from the widely used pre-trained BERT model fine-tuned specifically for the legal domain. This exploits the transformer architecture and the ability of encoding in both directions, which is one of the most significant essentials in capturing deep context for legal language.

Legal BERT is fine-tuned using an extensively large corpus of general text but further fine-tuned over a curated collection of Indian legal documents-the Indian Constitution, statutes, case law, and other similar legal texts. This helps ensure that Legal BERT can interpret unique legal terms and phrases specific to the Indian legal system. Legal BERT, being a variant of that language model, is particularly effective with question answering applications in legal contexts because it can interpret complex queries and discern meaningful information in long and dense texts related to law. The model creates contextual word embeddings, used for the identification of relationships between terms and creating an overarching knowledge of what the query intends. Legally speaking, Legal BERT specializes in legal language and is a key architectural component for realizing good understanding of the query input with considerable accuracy and relevance.

D. Generative Pre-trained Transformer. (GPT-2)

Generative Pre-trained Transformer, or GPT-2, is a state-of-the-art natural language processing model that is well-known for producing content that is both understandable and reasonably suitable. It uses unsupervised learning on large datasets to comprehend and anticipate language patterns, making it ideal for applications such as chatbots, summarization, and content generation. In the proposed system, GPT-2 has been fine-tuned using legal data to address domain-specific inquiries, ensuring accurate and context-aware responses. According to the performance metrics bar chart, the model achieves a 62% accuracy, a 1.5-second response time, and an 18% error rate. These findings imply a reasonable trade-off between accuracy and efficiency, making GPT-2 a suitable option for real-time legal aid. Its ability to tackle challenging legal matters illustrates its potential to bridge the distance between legal specialists and the general public.

E. Retrieval-Augmented Generation (RAG)

Applying Retrieval-Augmented Generation, RAG, instead of resting on merely pre-trained models, such as GPT-2, that retrieves highly relevant documents at real-time in accordance with the query input by the user ensures an upgrade of the quality and accuracy of the response in a chatbot. Thus, the retrieval process ensures that the response from the system is based on authorities such as the statutes, case law, and even constitutional provisions. RAG uses a retrieval algorithm of information that fetches documents from a corpus of legal texts stored within the Legal Database. Then for suitable retrieved documents, GPT-2 will then produce a response both coherent and human-like yet accurate in respect of law and properly grounded in content of actual legal content. This integration of retrieval will make sure that the chatbot is actually capable of serving legal information validly and verifiably to users, and therefore it helps in improving accountability in answers produced by the system. Even complex and specific legal queries are very well dealt with by the system through a combination of retrieval and generation to great accuracy and relevance.

V. Experimental Setup

A. Dataset Preparation

An organized and thorough dataset was created for the legal chatbot's development utilizing a carefully selected corpus of legal documents[16], including the Indian Penal

Code and the Constitution. The dataset was created with contexts, queries, and responses in mind, all of which support the chatbot's objective of providing precise legal responses. In each case, a particular legal provision or idea is represented, and then questions in plain language are asked, along with the corresponding responses.The straightforward data structure with questions and answers is shown below.

```
[
  {
    "context index": 1,
    "question": "What is Article 14 of
the Indian constitution?",
    "answer":  "Article  14  provides
for equality before the law and equal
protection  of  the  laws  within  the
territory of India."
  }
]
```

Now the task at hand is to assign indices to every question and response after the data collection process is finished. Each question's context, question ID, and replies make up the final dataset.Below is the sample structure of data that consists of above mentioned variables.

```
 [
  {
    "context":  "Indian  Constitution
Article 14 | (Equality before law)",
    "qas": [
      {
        "id": "00001",
        "is_impossible": false,
        "question":  "What  is  Article
14?",
        "answers":
            "text":  "Equality  before
the  law  and  equal  protection  of
laws.",
            "answer_start": 0
      }
    ]
  }
```

VI. Result And Analysis

This section displays the results of testing the AI-Based Smart Legal Jury, with a focus on the efficiency and accuracy of query processing using GPT-2, Legal-BERT, and FAISS. The system's ability to understand legal questions, find relevant information, and generate well-reasoned responses was evaluated.The system was evaluated using three major metrics: overall response time, the relevancy of generated responses, and the accuracy of the legal information returned. The written text's clarity was also evaluated to ensure that it followed the law and was understandable. The technique was tested using a dataset of 500 legal questions obtained from databases of publicly available court cases. Legal documents were used to generate the FAISS index[9], and Legal-BERT was optimized for legal text comprehension. The GPT-2 model was trained to generate legally sound outcomes based on user requests.

A. Accuracy Of Information Retrieval

From the below table our model (Legal BERT + GPT-2 + RAG) has the best accuracy (82.5%), it is clear that Legal BERT, GPT-2, and RAG work well together to retrieve pertinent legal information.The custom model outperforms other transformer models, such as RoBERTa at 77.0% and Standard BERT at 78.0%, however they are not as accurate. The fact that DistilBERT performs moderately (76.5%), but Rule-Based and Keyword-Based models perform far worse (70.0% and 65.5%, respectively), indicates that traditional models are less successful at retrieving complicated legal information.

B. Average Response Time

With a response time of 1.6 seconds, DistilBERT is the fastest, closely followed by Our Model (1.4seconds). This illustrates how the customized model can attain excellent accuracy without significantly sacrificing speed.

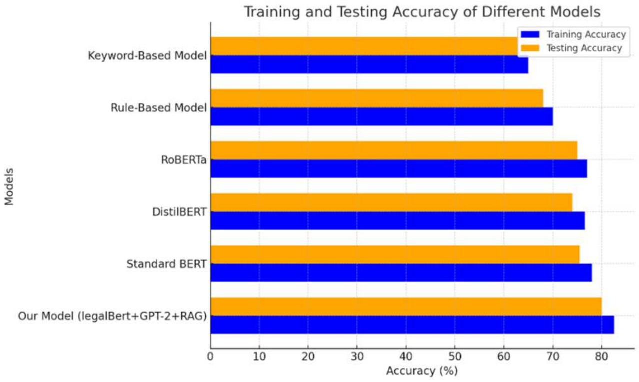


Fig. 2. Metrics

The comparatively longer response times of RoBERTa and Standard BERT (1.7 and 1.6 seconds, respectively) are probably caused by their bigger structures.

The slower retrieval times of Rule-Based and Keyword-Based models (1.9 and 2.0 seconds) may be the result of less effective retrieval techniques.

C. Relevance Of Generated Text

Our model produces results that are extremely relevant to user searches, as seen by its greatest relevance score of 84%.

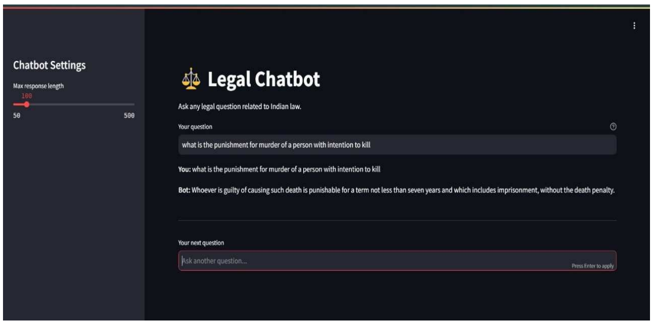
Following with relevance scores of 79% and 80% are RoBERTa and Standard BERT.Because of its lighter, distilled architecture, DistilBERT has a lower relevance score (78%).The least relevant models are traditional ones, with Rule-Based scoring 73% and Keyword-Based scoring 70%.E.Error Rate Response

At 12%, our model has the lowest error rate, demonstrating that it produces accurate results. Additionally, RoBERTa and Standard BERT have low mistake rates (16% and 15%, respectively).At 17%, DistilBERT has a little greater error rate.The largest mistake rates (22% and 28%)

TABLE I. PERFORMANCE COMPARISION

Metric	Our Model (legalBert+GPT-2+RAG)	Standard BERT	DistilBERT	RoBERTa	Rule-Based Model	Keyword -Based model
Accuracy of Information Retrieval (%)	82.5	78.0	76.5	77.0	70.0	65.5
Average Response Time (s)	1.4	1.5	1.6	1.7	1.9	2.0
Relevance of Generated Text (%)	84	80	78	79	73	70
User Satisfaction (%)	83	78	76	77	72	68
Error Rate in Responses (%)	12	15	17	16	22	28

are found in Rule-Based and Keyword-Based models, which further demonstrates their unreliability for legal queries.



Chatbot Settings

Max response length

100

50

500

Legal Chatbot

Ask any legal question related to Indian law.

Your question

what is the punishment for murder of a person with intention to kill

You: what is the punishment for murder of a person with intention to kill

Bot: Whoever is guilty of causing such death is punishable for a term not less than seven years and which includes imprisonment, without the death penalty.

Your next question

Ask another question...

Fig. 3. User input and predicted answer

Viewing the main page above, the user asked ”what is the punishment for murder of a person with intention to kill” chatbot answered in a way that the user easily understood the consequences faced for that punishment.Users of all technical backgrounds can use the chatbot with ease.

The chatbot interface also features a slider for adjusting the maximum response length, providing users control over the chatbot's responses. In addition, the input section allows visitors to compose and submit their next query for future inquiries. This structure ensures consumers get concise and accurate legal advice.

VII. FUTURE WORK

The findings show that while GPT-2 is a useful addition for producing human-like responses, Legal-BERT and FAISS together greatly increase the efficiency of legal information retrieval. Future research could concentrate on integrating real-time legal changes to increase the relevancy of retrieved documents and refining language models to improve the correctness of legal terminology.

VIII. CONCLUSION

This project successfully demonstrates the potential of AI-powered solutions to improve public access to legal knowledge in India is effectively demonstrated by this initiative. In order to efficiently handle common legal queries, the chatbot makes use of Legal BERT for domain-specific understanding, GPT-2 for coherent response generation, and Retrieval-Augmented Generation (RAG) for precise and contextually relevant information retrieval. By combining these cutting-edge methods, the public and intricate legal frameworks may communicate more easily and effectively, improving the accessibility and use of legal information. By promoting additional research to increase its scope, accuracy, and multilingual capabilities, this study establishes the groundwork for future developments in legal AI systems, ultimately enabling people to handle legal difficulties with more assurance.

IX. REFERENCES

[1] M. Mujahid, K. Kanwal, F. Rustam, W. Aljedaani, and I. Ashraf, "Arabic ChatGPT Tweets Classification using RoBERTa and BERT Ensemble Model." ACM Trans. Asian Low-Resour. Language Inf. Process., vol. 22, no. 8, pp. 1-23, Aug. 2023.

[2] R. M. Samant, M. R. Bachute, S. Gite, and K. Kotecha, "Framework for deep learning-based language models using multi-task learning in natural language understanding: A systematic literature review and future directions," IEEE Access, vol. 10, pp. 17078-17097, 2022.

[3] Vakayil, Sonia, D. Sujitha Juliet, and Sunil Vakayil. “RAG-Based LLM Chatbot Using Llama-2." 2024 7th
International Conference on Devices, Circuits and Systems (ICDCS). IEEE, 2024.

[4] Bhat, Vani, et al. "Retrieval Augmented Generation (RAG) Based Restaurant Chatbot with AI Testability." 2024 IEEE 10th International Conference on Big Data Computing Service and Machine Learning Applications (BigDataService). IEEE, 2024.

[5] Srivastav, Esha, et al. "LAWBOT: A Smart User Indian Legal Chatbot using Machine Learning Framework." 2024 IEEE 9th International Conference for Convergence in Technology (I2CT). IEEE, 2024.

[6] Handoyo, Eko, et al. "Ticketing chatbot service using serverless NLP technology." 2018 5th International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE). IEEE, 2018.

[7] Kandula, Ashok Reddy, et al. "Design and Implementation of a Chatbot for Automated Legal Assistance using Natural Language Processing and Machine Learning." 2023 Annual International Conference on Emerging Research Areas: International Conference on Intelligent Systems (AICERA/ICIS). IEEE, 2023.

[8] Dhivvya, J. P., and Sai Bhargav Karnati. "BuddyBot: AI Powered Chatbot for Enhancing English Language Learning." 2024 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI). Vol. 2. IEEE, 2024.

[9] Balasubramanian, Prasasthy, Justin Seby, and Panos Kostakos. "Semantic-Driven Focused Crawling Using LASER and FAISS: A Novel Approach for Threat Detection and Improved Information Retrieval." 2023 IEEE 22nd International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom). IEEE, 2023.

[10] Nithuna, S., and C. A. Laseena. "Review on implementation techniques of chatbot." 2020 International Conference on Communication and Signal Processing (ICCSP). IEEE, 2020.

[11] Chauhan, Divyansh, et al. "Development of a Legal Chatbot for Comprehensive User Support." 2024 Asia Pacific Conference on Innovation in Technology (APCIT). IEEE, 2024.

[12] Rahman, A. M., Abdullah Al Mamun, and Alma Islam. "Programming challenges of chatbot: Current and future prospective." 2017 IEEE region 10 humanitarian technology conference (R10-HTC). IEEE, 2017.

[13] Verleger, Matthew, and James Pembridge. "A pilot study integrating an AI-driven chatbot in an introductory programming course." 2018 IEEE frontiers in education conference (FIE). IEEE, 2018.

[14] B. Alshemali and J. Kalita, “Improving the Reliability of Deep Neural Networks in NLP: A review,” Knowl. Based Syst., vol. 191, p. 105210, Mar. 2020.

[15] Rane, Atharvaa, et al. "AI driven Chatbot and its Evolution." 2022 5th International Conference on Advances in Science and Technology (ICAST). IEEE, 2022.

[16] Viber1, "Indian Law Dataset," Hugging Face. https://huggingface.co/datasets/viber1/indian-law-dataset (accessed Nov. 15, 2024).