PhishBENCH: A benchmark for legal tasks related to phishing attacks

Ioannis Katoikos* *i.katoikos@ac.upatras.gr*
Dimitrios Kosmopoulos† *dkosmo@upatras.gr*
Ioannis Iglezakis ‡ *iingleza@law.auth.gr*
Panagiota Kiortsi‡ *kiortsip@gmail.com*
George A. Thomopoulos* *gthomop@upatras.gr*
Christos A. Fidas* *fidas@upatras.gr*

***Abstract*—Phishing attacks constitute a significant cyber-threat, causing financial losses and undermining digital trust. Despite advancements in awareness campaigns and regulatory efforts, many individuals and organizations struggle to understand their legal rights and obligations when victimized by phishing. This paper presents a benchmark framework designed to support researchers in the analysis of legal texts and cases, with the goal of helping users navigate the complex legal landscape surrounding phishing attacks. We evaluate BERT-based models, large language models, and retrieval-augmented generation using a curated knowledge base of phishing laws, regulations, and common scenarios. We provide baseline evaluations across multiple task categories, including true-false questions, phishing attack cases, penal code analysis, legal cases with judicial decisions, and question-answering tasks, all within the Greek legal jurisdiction.**

***Index Terms*—benchmark, phishing law, large language model, retrieval augmented generation**

I. Introduction

Phishing is a type of cyber-attack that involves tricking individuals into disclosing personal, financial, or other sensitive information or unwittingly downloading malicious software. This is typically achieved through deceptive communications, often email messages (but also other communication channels), that mimic legitimate organizations or individuals.

Phishing attacks have recently become increasingly dangerous and sophisticated, often leveraging deepfake technology—particularly voice impersonation—to deceive victims e.g., [7]. The situation is further worsened by Phishing-as-a-Service (PhaaS), a growing cybercrime model that allows even non-technical users to carry out sophisticated phishing attacks using subscription-based toolkits and services available on underground marketplaces [21].

For individuals, falling victim to a phishing attack can lead to immediate financial loss, as attackers may gain access to bank accounts, credit cards, or online payment platforms. Beyond the financial damage, individuals may lose access to important personal files or accounts, especially when phishing leads to ransomware infections or the compromise of cloud storage services. For organizations, the stakes are even higher. These breaches often come with heavy financial consequences, from direct theft and ransom payments to the significant costs of incident response, legal action, regulatory fines and reputational damage, see e.g., [2], [3]. The FBI has received over 4 million complaints related to internet scams in the past five years, with estimated losses exceeding 50 billion USD and impacting individuals worldwide [16].

These challenges highlight the need for close co-operation with legal experts who are well-versed in relevant laws, enabling swift action and helping to avoid additional costs. However, legal advice is not always readily available when urgently needed. In such cases, digital assistants may help bridge this gap by providing timely support and guidance.

In our research, we aim to develop automated systems for providing first-aid legal advice to individuals and organizations affected by phishing attacks. We contribute by introducing:

- A comprehensive benchmark comprising multiple task types—true-false questions, phishing attack case analysis, penal code interpretation, legal decision prediction and Q&A—designed to evaluate performance on phishing-related law within Greek jurisdiction.
- Comparative evaluations of BERT-based models and Large Language Models (LLMs), including retrieval-augmented generation (RAG) systems.

This work has been financially supported by the Hellenic Foundation for Research & Innovation (HFRI) under the project entitled "Artificial Intelligence-driven Framework and Legal Advice Tools for Phishing Prevention and Mitigation in Information System" (AILA) with Proposal ID 15440.

*Department of Electrical and Computer Engineering, *University of Patras*, Greece
†Department of Computer Engineering and Informatics, *University of Patras*, Greece
‡Faculty of Law, *Aristotle University of Thessaloniki*, Greece

2375-0197/25/$31.00 ©2025 IEEE
DOI 10.1109/ICTAI66417.2025.00198

II. Related works

Benchmarks: LegalBench represents an ambitious collaborative effort in legal AI evaluation. This benchmark consists of 162 tasks gathered from 40 contributors, covering six different types of legal reasoning through a unique crowd-sourcing effort within the legal community [11]. LexGLUE (Legal General Language Understanding Evaluation) provides standardized evaluation across diverse legal NLP tasks. This benchmark offers a collection of datasets for evaluating model performance across legal tasks in a standardized way, inspired by the success of GLUE in general NLP [5]. CLERC (Case Law Evaluation and Retrieval Corpus) focuses on legal information retrieval and retrieval-augmented generation. Built upon the Caselaw Access Project containing over 1.84 million federal case documents, it evaluates models' ability to find legal citations and generate coherent legal analyses [15]. CaseHOLD (Case Holdings On Legal Decisions) presents a fundamental legal reasoning task through multiple-choice questions. It contains over 53,000 questions designed to identify relevant case holdings, which represent the governing legal rules applied to specific factual scenarios. [27]. The dataset was created from 3,446,187 legal decisions across all federal and state courts, representing approximately 37GB of text. Privacy Law Datasets focus on the growing importance of privacy regulation and policy analysis. The Privacy Law Corpus includes 1,043 privacy laws covering 182 jurisdictions worldwide [12], while PrivacyQA provides 1,750 questions about privacy policies with expert annotations [22].

Assistants on phishing issues: AI-based chatbots can play a role in phishing from both perspectives—serving either the attacker or the defender. In [14] a chatbot is used to educate users on the identification of phishing emails. In [26] a chatbot is developed for the identification of phishing in social networks. In [23] LLMs are used to design phishing attacks and a BERT-based tool is developed to counter those attacks. In [24] LLMs are used to design attacks based on SMS messages (smishing). In [6] GPT3.5, GPT4 and a customized such model were used to classify email messages to phishing-non phishing ones.

Despite numerous efforts in related areas, we are not aware of any similar initiative focused on developing benchmarks or assistants for legal tasks on phishing. Our work aims to address this research gap by introducing a benchmark for six tasks and by providing baselines using NLP and LLM models. Our work is expected to pave the way for smart assistants on legal phishing issues.

III. The methods

The models to evaluate against PhishBENCH include BERT and LLM - based models.

A. BERT-based models

BERT [8] employs transformer architecture and is trained using Masked Language Modeling (MLM) to predict masked tokens from context and Next Sentence Prediction (NSP) to learn inter-sentence relationships. Several BERT variants of have been developed for specific tasks:

1) BERT-Base-Uncased: It is the original BERT release with 12 transformer layers, 768 hidden units, and 110M parameters [8]. The "uncased" version converts text to lowercase and removes accents, reducing vocabulary size. It was pre-trained on BookCorpus (800M words) and English Wikipedia (2.5B words), becoming an NLP benchmark due to strong performance and manageable computational requirements.

2) RoBERTa-Base: It is is Facebook AI's optimized BERT reimplementation with similar architecture (12 layers, 768 hidden units, 125M parameters) but trained on 10x more data, using dynamic masking, removing NSP, and employing longer sequences and larger batches [20]. These optimizations improve performance on NLP benchmarks like GLUE and SQuAD. Unlike BERT, RoBERTa is case-sensitive.

3) DistilBERT: DistilBERT is a compressed BERT model created through knowledge distillation, achieving 40% fewer parameters (66M), 60% faster inference, and 97% of BERT's GLUE performance. It reduces encoder layers from 12 to 6 while maintaining BERT's vocabulary and tokenization, making it suitable for resource-constrained environments.

4) Legal-BERT: Legal-BERT is a domain-specific BERT adaptation developed by NCSR "Democritos," further pre-trained on EU legislation, court rulings, and English legal texts [4]. It outperforms general models on legal tasks like document classification, entailment, and information extraction by better capturing legal terminology and discourse structure.

B. Large Language Models (LLMs)

Large Language Models have transformed NLP from text encoding tasks (e.g., BERT) to text generation, reasoning, and dialogue applications. These decoder-based Transformers are causally trained on web-scale data with zero-shot, few-shot, and instruction-following capabilities. Mistral [17] represents a promising open-source LLM family with permissive licensing for academic and commercial use.

We evaluated Mistral-7B (unsloth/mistral-7b-instruct-v0.3) on legal classification and article extraction, leveraging its prompt-driven capabilities for minimal fine-tuning prototyping. Key features include sliding window attention, 32K token contexts via RoPE scaling, compact architecture for local deployment, and instruction-following for rapid task prototyping.

Mistral demonstrates strong multilingual capabilities, particularly in Greek comprehension, effectively handling Greek prompts, documents, and labels with few-shot examples. It tolerates mixed-language inputs (Greek legal text with English instructions), common in Greek legal NLP. While not adapted on Greek legal corpora, its multilingual web training provides satisfactory performance with well-crafted prompts.

To enhance LLM performance and generate more accurate, context-specific answers, we used Retrieval-Augmented Generation (RAG) architecture. Context is retrieved based on user queries and incorporated into the LLM prompt. We segmented the corpus into documents, processed them with OpenAI's text-embedding-3-small, and stored embeddings in a vector index. During inference, queries are embedded and retrieve top 10 documents through hybrid retrieval (embedding similarity and keywords), which are then fed into the LLM prompt for answer generation.

IV. The Dataset

We constructed a comprehensive legal-phishing benchmark from diverse sources including legal documents, phishing scenarios [10], GDPR [9], Greek Penal Code [13], and relevant Greek court cases and decisions. A dedicated team of legal experts specializing in phishing authored principal queries and ground-truth answers. To expand our dataset, we further leverage the available source documents in combination with structured prompts and generations from OpenAI's API. All generated queries were human-evaluated by legal experts and refined to ensure legal and contextual accuracy. The final dataset [19] is publicly available and comprises of:

a. True/False questions We assembled 869 True/False questions related to the Greek Penal Code, GDPR phishing provisions, and general phishing scenarios, with 396 False and 473 True labels.

b. Phishing attack cases This task classifies unseen phishing cases into predefined scenario types based on legal and cybercrime contexts. The dataset contains 1252 queries across seven phishing attack types, including generated and real-world examples (credential harvesting, financial fraud, impersonation). The distribution across categories is displayed is Table I.

TABLE I
Number of queries per phishing type

Phishing Type	Number of Queries
Email Phishing / Smishing	236
Man-in-the-Middle Attack	164
Quishing (QR Code Phishing)	169
SIM Swapping	130
Social Media Phishing	201
Spear Phishing	183
Vishing (Voice Phishing)	169

c. Penal Code This task classifies questions by Greek cybersecurity penal code articles using 1,917 queries across 11 labels. The challenge lies in high article similarity and overlapping legal terminology. Table II shows the distribution of the dataset.

TABLE II
Number of queries per article of the Greek penal code

Article Reference	Queries
Approaching child (Art. 348B)	175
Child pornography (Art. 348A)	183
Computer fraud (Art. 386A ΠΚ)	200
Genital respectability (Art. 337)	191
Personal data protection (Art. 22 N. 2472/1997)	149
Rights management info (Art. 66B N. 2121/1993)	194
Technological measures (Art. 66A N. 2121/1993)	164
Unauthorized data access (Art. 370Γ ΠΚ)	152
Use of software (Art. 370Γ ΠΚ)	180
Use of software (Art. 66 N. 2121/1993)	179
Violation of secrecy (Art. 370B ΠΚ)	150

d. Legal Cases This task classifies questions and scenarios by specific Greek legal cases. Each instance corresponds to a case label from Greek judicial decisions, phishing-related court rulings, and generated cases based on real legal documents. The dataset contains 1,689 queries across 15 legal case labels. Table III shows query distribution.

TABLE III
Number of generated queries per legal case

Legal Case	Queries
245/2023 Chalkida First Instance	166
Athens Appeal 387/2013	126
Athens Magistrate 1496/2024	80
Rhodes Misdemeanor 21/2022	120
Athens Appeal 1312/2018	113
Athens Appeal 678/2013	74
Mytilene Magistrate 24/2023	123
Thessaloniki Peace 1772/2023	117
Athens First Instance 546/2024	152
Rodopi First Instance 4/2023	103
Supreme Court 1414/017	107
Thessaloniki First Instance 1894/2023	118
Thessaloniki First Instance 2297/2022	109
Thessaloniki First Instance 232/203	81
Thessaloniki Peace 2015/2022	100

e. Guilty or Innocent This binary classification task predicts guilty/not guilty verdicts for legal scenarios inspired by Greek cases (see e.g., [25] for similar tasks). The challenge requires implicit judgment and interpretive logic, where similar facts may yield different outcomes due to legal nuances. The dataset contains 1,216 queries (632 guilty, 584 innocent).

f. Q&A This generative task answers legal questions using Greek Penal Code contexts through 2,260 manually and programmatically generated question-answer pairs. We evaluate BERT models on information extraction effectiveness and LLMs (particularly Mistral) on generating semantically similar answers to ground

truth while avoiding off-topic responses. BERT required SQuAD preprocessing with context and answer start/end indices.

V. Experimental evaluation

A. BERT-Based Model setup

Data was split 80/20 for training/testing, with BERT variants optimized using Optuna [1] hyperparameter tuning. Optuna uses advanced search algorithms like TPE and random sampling, more efficient than grid search, with trial pruning to terminate unpromising configurations early. Each trial employed 5-fold stratified cross-validation with early stopping after two consecutive epochs without validation loss improvement. Key optimized hyperparameters included *per device train batch size*: 4-16 (balancing memory and stability), *learning rate*: 1e-5 to 5e-5 (optimal convergence), *weight decay*: 0.1-0.3 (preventing overfitting), and *number of epochs*: 2-10 (adequate training without overfitting).

For the tasks (a-e) we computed mean F1-scores across folds, conducting 20 trials and selecting the best-performing configuration for final training. The model was retrained on the entire training set using the best configuration, with the median number of epochs from the best trial (due to early stopping variations across folds). Final evaluation on the test set reported accuracy, precision, recall, and F1-score.

For the task (f), we used two metrics: *Exact Matching (EM)*: Percentage of predictions exactly matching ground truth after normalization (lowercasing, removing punctuation). Awards credit only for fully correct predictions. *F1-score*: Harmonic mean of token-level precision and recall, allowing partial credit for overlapping predictions.

B. LLM-Based Model setup

For the Large Language Model approach we focused on three independent strategies involving:

1) Supervised Fine-Tuning: We fine-tuned Mistral using LoRA-based parameter-efficient fine-tuning on our datasets through supervised learning. Following the same procedure as BERT evaluation, we used Optuna over 20 trials with 5-fold stratified cross-validation and early stopping. Optimized parameters included batch size, learning rate, weight decay, and epochs (same ranges as BERT). The best configuration was retrained on the full training set and evaluated on the test set.

2) Further Pretraining via Causal Language Modeling: We conducted unsupervised continued pretraining using Causal Language Modeling (CLM) to enrich Mistral's legal and phishing domain knowledge. CLM predicts the next token based on preceding tokens in a left-to-right manner, unlike MLM which uses surrounding context. The pretraining corpus used the same documents as our dataset, formatted as (title, body) pairs to help the model learn domain-specific syntax, vocabulary, and discourse patterns without supervised labels.

We used again Optuna to optimize pretraining hyperparameters including learning rate, weight decay, and epochs. Multiple trials with random sampling evaluated performance based on training stability and loss convergence on validation segments. The chosen configuration consistently produced stable training and efficient convergence: *Device train batch size*: 4, *Gradient accumulation steps*: 4, *Number of training epochs*: 10, *Warmup ratio*: 0.05, *Learning rate*: 5e-5, *Embedding learning rate*: 1e-5, *Weight decay*: 0.01.

3) Retrieval-Augmented Generation: Since LLMs are generative rather than encoding models, we constructed prompts to enable classification functionality. These prompts outlined specific tasks and included few-shot examples for clarity. Prompt details are in the Appendix.

Evaluation for tasks (a-e) used accuracy, precision, recall, and F1-score metrics. For the task (f), we utilized two metrics:

Sentence Similarity: Measures semantic closeness between ground truth and LLM predictions using pretrained embeddings. The embedding model converts each sentence into fixed-length dense vectors representing meaning, then computes cosine similarity. This captures semantic meaning rather than exact word matches. We used Sentence-T5, a top performer for semantic similarity on long, nuanced generative outputs.

BERT Score: This metric evaluates token-level semantic similarity between predictions and ground truth by aligning their BERT embeddings. Both sentences are tokenized, and each token receives contextualized embeddings. For each candidate token, the most similar reference token is identified via cosine similarity (and vice versa). Precision, recall, and F1 are computed from these alignments. *Precision*: Measures how well each token in the predicted text is matched by tokens in the ground truth. *Recall*: Measures how well each token in the ground truth is matched by tokens in the predicted text. *F1 Score*: A harmonic mean of precision and recall.

We emphasize the recall metric, as it measures how well LLM predictions represent the ground truth—critical information for our evaluation.

C. Experimental Results

1) True - false: Surprisingly, domain-specific models like Legal-BERT slightly underperformed compared to general-purpose models, suggesting generalization may outweigh specialization (Table IV). RAG achieved the best zero-shot performance (F1: 0.816), while our fine-tuned model achieved the highest overall F1 score (0.833), demonstrating domain-specific fine-tuning advantages. Classical transformers (BERT-large, RoBERTa-base) remained competitive.

TABLE IV
PERFORMANCE OF VARIOUS MODELS ON THE TRUE/FALSE LEGAL CLASSIFICATION TASK.

Model	Best Accuracy	Best F1	Final Accuracy	Final F1
bert-base-uncased	0.781	0.781	0.753	0.752
FacebookAI/roberta-base	**0.794**	**0.794**	**0.776**	**0.772**
nlpaueb/legal-bert-base-uncased	0.747	0.714	0.736	0.735
distilbert-base-uncased	0.78	0.779	0.741	0.741
Llama Index (RAG)	–	–	**0.816**	**0.815**
Mistral Pretraining baseline	–	–	0.736	0.723
Fine-tuned Mistral	**0.803**	**0.802**	**0.834**	**0.833**

TABLE V
PERFORMANCE OF VARIOUS MODELS ON THE PHISHING SCENARIOS CLASSIFICATION TASK. ALL TRANSFORMER MODELS WERE FINE-TUNED.

Model	Best Accuracy	Best F1	Final Accuracy	Final F1
Facebook/roberta-base	0.925	0.925	0.9	0.897
bert-base-uncased	0.919	0.919	0.888	0.886
nlpaueb/legal-bert-base-uncased	**0.917**	**0.917**	**0.912**	**0.911**
distilbert-base-uncased	0.874	0.873	0.837	0.834
Llama Index (RAG)	–	–	0.574	0.544
Mistral Pretraining baseline	–	–	0.594	0.546
Fine-tuned Mistral	0.641	0.608	0.661	0.635

2) Phishing scenarios: Legal-BERT achieved exceptional performance, outperforming all models with F1-score 0.911 (Table V). BERT-base-uncased and RoBERTa showed competitive results, suggesting legal domain pretraining advantages for phishing classification. RAG and pretrained-only models exhibited limited generalization to novel phishing cases, highlighting the need for specific training. The fine-tuned model achieved moderate success (F1: 0.635), indicating room for improvement through hyperparameter optimization or data augmentation.

3) Greek penal code articles: All fine-tuned transformers showed strong performance, with RoBERTa achieving the highest validation F1-score of 0.982 (Table VI). Legal-BERT demonstrated competitive performance understanding legal context but didn't outperform generalized models, which may benefit from broader training data. RAG methods and pretrained-only models struggled with this task, lacking necessary fine-tuning for legal nuances. Fine-tuned models faced generalization challenges, possibly due to overfitting or insufficient ability to differentiate between similar articles.

4) Phishing legal cases: All transformer models demonstrated near-perfect classification performance, indicating strong task learnability with sufficient capacity and domain-specific fine-tuning (Table VII). Despite this, baseline and RAG models performed poorly. The fine-tuned LLM, trained with fewer epochs and limited context, struggled to generalize (test F1: 0.064). DistilBERT achieved almost 100% F1-score despite being smaller, confirming reliable mapping. RAG and pretrained-only models lacked context awareness for multi-label legal classification, underscoring the importance of task-specific fine-tuning in legal NLP.

5) Guilty or Innocent: Fine-tuned transformers significantly outperformed baseline and RAG approaches (Table VIII). DistilBERT achieved the highest accuracy and F1-score despite being lightweight, while BERT and RoBERTa outperformed Legal-BERT. RAG with Llama Index showed considerably lower performance, indicating generative approaches are suboptimal for classification without customized prompts and retrievers. Pretraining-only baselines performed poorly, highlighting fine-tuning importance. Mistral fine-tuning showed no improvement for this task.

6) Q&A: BERT-based models showed strong performance in legal question answering with appropriate context. Classic BERT achieved highest scores (EM: 0.759, F1: 0.908), indicating high precision and ground truth alignment. Legal-BERT performed competitively but didn't surpass general BERT, suggesting domain-specific pretraining may not always enhance performance when tasks are well-addressed by general patterns and fine-tuning. RoBERTa and DistilBERT yielded commendable results (F1 $>$ 0.85), demonstrating potential for efficient deployment despite slightly lower EM scores.

Fine-tuned Mistral achieved best performance (recall: 0.759, semantic similarity: 0.872), demonstrating superior content coverage and semantic alignment. RAG with Mistral yielded competitive results (recall: 0.753, similarity: 0.848), showing strong semantic alignment despite slightly lower recall. Pretraining-only Mistral underperformed in both metrics, highlighting that domain pretraining alone is insufficient without task-specific fine-tuning. Higher semantic similarity scores correlate with better recall, reinforcing that semantic alignment is

TABLE VI
PERFORMANCE OF VARIOUS MODELS ON THE ARTICLE CLASSIFICATION TASK. ALL TRANSFORMER MODELS WERE FINE-TUNED.

Model	Best Accuracy	Best F1	Final Accuracy	Final F1
bert-base-uncased	0.984	0.984	0.979	0.979
Facebook/roberta-base	**0.987**	**0.987**	**0.982**	**0.982**
nlpaueb/legal-bert-base-uncased	0.984	0.984	0.979	0.979
distilbert-base-uncased	0.98	0.98	0.977	0.977
Llama Index (RAG)	–	–	0.621	0.617
Mistral Pretraining baseline	–	–	0.568	0.55
Fine-tuned Mistral	0.665	0.637	0.659	0.625

TABLE VII
PERFORMANCE OF VARIOUS MODELS ON THE CASE CLASSIFICATION TASK. ALL TRANSFORMER MODELS WERE FINE-TUNED. FINAL SCORES REFER TO VALIDATION OR TEST PERFORMANCE DEPENDING ON AVAILABILITY.

Model	Best Accuracy	Best F1	Final Accuracy	Final F1
bert-base-uncased	0.996	0.996	1.000	1.000
Facebook/roberta-base	1.000	1.000	1.000	1.000
nlpaueb/legal-bert-base-uncased	1.000	1.000	1.000	1.000
distilbert-base-uncased	0.998	0.998	1.000	1.000
Llama Index (RAG)	–	–	0.056	0.031
Mistral Pretraining baseline	–	–	0.083	0.024
Fine-tuned Mistral	0.02072	0.01395	0.062	0.064

TABLE VIII
PERFORMANCE OF VARIOUS MODELS ON THE GUILTY/NOT GUILTY CLASSIFICATION TASK. ALL TRANSFORMER MODELS WERE FINE-TUNED.

Model	Best Accuracy	Best F1	Final Accuracy	Final F1
Facebook/roberta-base	0.889	0.889	0.893	0.893
bert-base-uncased	0.882	0.882	0.898	0.898
nlpaueb/legal-bert-base-uncased	0.879	0.877	0.885	0.884
distilbert-base-uncased	**0.89**	**0.89**	**0.93**	**0.93**
Llama Index (RAG)	–	–	0.582	0.565
Mistral Pretraining baseline	–	–	0.627	0.603
Fine-tuned Mistral	0.558	0.713	0.607	0.59

TABLE IX
PERFORMANCE OF VARIOUS BERT MODELS ON THE Q&A TASK.

Model	Best EM score	Best F1	Final EM score	Final F1
Facebook/roberta-base	0.696	0.832	0.719	0.854
bert-base-uncased	0.736	0.883	0.759	0.908
nlpaueb/legal-bert-base-uncased	0.717	0.858	0.708	0.863
distilbert-base-uncased	0.698	0.848	0.706	0.874

essential for meaningful legal content generation.

Overview: Fine-tuned BERT-based transformers consistently outperformed alternatives across legal tasks, with RoBERTa-base excelling in phishing detection and guilt assessment. While Legal-BERT showed promise, general-purpose models often matched or exceeded domain-specific performance, suggesting broader pretraining benefits. RAG demonstrated strong zero-shot capabilities for simple tasks but struggled with complex reasoning requiring deeper contextual understanding. Models relying solely on pretraining showed limited generalization, highlighting fine-tuning's critical importance in legal NLP. LLMs achieved competitive results through RAG (combining retrieval with generation) and direct fine-tuning, demonstrating versatility across context-rich and context-agnostic scenarios. Hyperparameter optimization using frameworks like Optuna proved essential for optimal performance. Fine-tuning on domain-specific data maximizes legal NLP performance, with even lightweight models like DistilBERT approaching state-of-the-art results when properly tuned. BERT variants excel at semantic query understanding, making them valuable for RAG preprocessing to identify user intent and enable more accurate document filtering in legal applications.

VI. DISCUSSION

This study focused on cybersecurity and phishing within the Greek Penal Code and European Union law,

TABLE X
PERFORMANCE OF MISTRAL ON DIFFERENT TRAINING STRATEGIES ON THE Q&A TASK.

Model	Best Recall	Best Sentence Similarity Score	Final Recall	Final Sentence Similarity Score
Llama Index (RAG)	–	–	0.753	0.848
Mistral Pretraining baseline	–	–	0.657	0.804
Fine-tuned Mistral	0.76	0.861	0.759	0.872

establishing a comprehensive benchmark within Greek jurisdiction and providing six legal tasks which can be extended to other jurisdictions and entities like the EU or national entities, given the shared nature of legal principles.

Nevertheless, domain adaptation remains challenging in cybersecurity, where concepts are highly specialized and underrepresented in general pretraining. Legal frameworks are fragmented across data protection and criminal law domains, lacking centralized codes, necessitating careful curation for comprehensive jurisdiction-specific datasets.

Our findings suggest that BERT models can standardize task labels. This can improve RAG architectures for efficient document searches across through article mapping, legal entity recognition, and case classification. Future research could explore cross-country comparative NLP to achieve harmonized classification of legal texts across jurisdictions. While knowledge transfer shows promise for enhancing performance across jurisdictions [18], comprehensive experimental validation remains an open research question that we plan to address in future work.

We primarily evaluated the Mistral model based on metrics such as recall, semantic similarity, and retrieval accuracy, standard in large language model (LLM) generation evaluations. However, we recognize that these metrics do not capture essential elements like factual consistency, hallucination rates, or potential biases in model outputs. This is particularly significant in the legal domain, where precision in facts, transparency in reasoning, and neutrality are imperative. This challenge is common in any attempt at domain adaptation via LLMs. Mistral may produce plausible but factually incorrect or irrelevant outputs. In high-stakes scenarios, such behavior could lead to misleading outputs or erroneous legal inferences. Furthermore, the black-box nature of LLMs hampers interpretability, making it difficult to audit or trace the reasoning behind decisions, which is a critical requirement in regulated fields like law. In future work, we plan to complement quantitative metrics with qualitative assessments, such as manual error analysis, hallucination audits, and bias detection across various legal sub-domains. Integrating structured evaluation frameworks (e.g., TruthfulQA, LegalBench) and human-in-the-loop validation will be necessary to ensure that legal LLM systems are both effective and trustworthy.

Despite near-perfect F1 scores, particularly in the legal case classification task (experiment d), suggesting strong ability to capture relevant patterns, there are still concerns. The task design may not have been sufficiently challenging, raising the risk of overfitting and limiting confidence in the model's generalization ability. We acknowledge that these results may partially reflect model memorization. This risk is heightened when the evaluation is conducted on test splits that are structurally similar to the training dataset. To mitigate this issue, future evaluations will include more adversarial examples created by legal professionals to better assess the model's ability to generalize beyond narrow legal task boundaries. Another approach could involve removing semantically similar queries and introducing noise to test robustness under more realistic and imperfect conditions, thus simulating what would occur in real-world legal NLP deployments.

Another limitation lies in the nature of the data. Our tasks rely on AI-generated legal questions and answers derived from real legal documents and reviewed by experts. While this enables large-scale experimentation across diverse legal scenarios, it still may reduce exposure to authentic legal language, adversarial queries, and domain-specific ambiguities, limiting real-world applicability. Nonetheless, the dataset offers valuable insights, such as contextualizing user queries and aligning responses with ground truth. Future work should benchmark against more user questions, court transcripts, or legal chat logs, incorporating adversarial prompting and human-authored datasets to better assess robustness, reliability, and risk of misinterpretation under realistic deployment conditions.

Finally, the ethical and legal implications of this work must be acknowledged. AI systems in the legal domain, while powerful, lack true legal reasoning, accountability, and the capacity to deliver reliable legal advice. Their use should therefore remain cautious: they may support information retrieval and document analysis but cannot replace qualified professionals. Future research should focus on risk mitigation through human-in-the-loop validation, transparency, and alignment with legal ethics, while also addressing regional compliance with data protection, liability, and professional standards before

practical deployment.

VII. Conclusion

This work presents a domain-specific legal NLP benchmark centered in Greek and European cybersecurity law, offering both classification and generative tasks groundend in legal documents. Through these tasks, we demonstrate the practical utility of fine-tuned open-source transformer models for structured legal language and context understanding, despite their limitations in factuality, transparency and generalization.

We chose to focus on open models due to their licensing freedom, local deployment, and reproducibility, which are essential for legal AI research. While we acknowledge that stronger closed-source models such as GPT-4, Claude or Gemini may offer state-of-the-art performance with low cost of finetuning, their restricted access and opacity limit their use in high-stakes domains. Future work will include benchmarking these models alongside open alternatives and further explore domain adaptations through jurisdictions.

Our benchmark offers insights into practical legal NLP workflows, including few-shot learning, classification, and prompt-based reasoning. These insights lay the foundation for more robust, multilingual, and trustworthy legal NLP systems. Ultimately, this work contributes to the development of responsible, reproducible, and transparent legal AI pipelines that can support, and not replace, qualified legal professionals.

References

[1] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama. Optuna: A next-generation hyperparameter optimization framework. In *ACM SIGKDD*, 2019.

[2] N. T. Aleke et al. Nation-state cyber attacks on critical infrastructure: A case study and analysis of the 2014 sony pictures hack by north korea. In *Leveraging Large Language Models for Quantum-Aware Cybersecurity*, pages 143–168. 2025.

[3] J. Beerman, D. Berent, Z. Falter, and S. Bhunia. A review of colonial pipeline ransomware attack. In *IEEE/ACM CCGridW*, pages 8–15, 2023.

[4] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos. LEGAL-BERT: The muppets straight out of law school. In T. Cohn, Y. He, and Y. Liu, editors, *EMNLP*, pages 2898–2904, Nov. 2020.

[5] I. Chalkidis, A. Jana, D. Hartung, M. Bommarito, I. Androutsopoulos, D. Katz, and N. Aletras. LexGLUE: A benchmark dataset for legal language understanding in English. In *ACL*, pages 4310–4330, May 2022.

[6] R. Chataut, P. K. Gyawali, and Y. Usman. Can ai keep you safe? a study of large language models for phishing detection. In *IEEE CCWC*, pages 0548–0554, 2024.

[7] A. de Rancourt-Raymond and N. Smaili. The unethical use of deepfakes. *Journal of Financial Crime*, 30(4):1066–1077, 2023.

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pages 4171–4186, 2019.

[9] European Parliament and Council. Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (general data protection regulation), 2016. OJ L 119, 4.5.2016, p. 1–88.

[10] Fortinet. 19 types of phishing attacks with examples. https://www.fortinet.com/resources/cyberglossary/types-of-phishing-attacks, 2025. Accessed July 2025.

[11] N. Guha and et al. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models, 2023.

[12] S. Gupta, G. Gopi, H. Balaji, E. Poplavska, N. O'Toole, S. Arora, T. Norton, N. Sadeh, and S. Wilson. Creation and analysis of an international corpus of privacy laws. *arXiv preprint arXiv:2206.14169*, 2022.

[13] Hellenic Penal Code. Greek penal code, 2024. As amended by Law 4577/2018.

[14] S. Hobert, A. Følstad, and E. L.-C. Law. Chatbots for active learning: A case of phishing email identification. *International Journal of Human-Computer Studies*, 179:103108, 2023.

[15] A. B. Hou, O. Weller, G. Qin, E. Yang, D. Lawrie, N. Holzenberger, A. Blair-Stanek, and B. V. Durme. Clerc: A dataset for legal case retrieval and retrieval-augmented analysis generation. *arXiv preprint arXiv:2406.17186*, 2024.

[16] Internet Crime Complaint Center, Federal Bureau of Investigation. 2024 internet crime report. Technical report, IC3, 2024. Accessed: 2025-06-21.

[17] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

[18] Z. Kang, H. Cai, X. Ji, J. Li, and N. Gu. Jurisctc: Enhancing legal judgment prediction via cross-domain transfer and contrastive learning, 2025.

[19] I. Katoikos, D. Kosmopoulos, I. INGLEZAKIS, P. Kiortsi, G. Thomopoulos, and C. Fidas. Phishbench dataset, Sept. 2025.

[20] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[21] J. Osamor, M. Ashawa, A. Shahrabi, A. Philip, and C. Iwendi. The evolution of phishing and future directions: A review. In *ICCWS*, volume 20, 2025.

[22] A. Ravichander, A. W. Black, S. Wilson, T. Norton, and N. Sadeh. Question answering for privacy policies: Combining computational and legal perspectives. In *EMNLP-IJCNLP*, pages 4947–4958, Nov. 2019.

[23] S. S. Roy, P. Thota, K. V. Naragam, and S. Nilizadeh. From chatbots to phishbots?: Phishing scam generation in commercial large language models. In *IEEE Symp. on Security and Privacy (SP)*, pages 36–54, 2024.

[24] A. M. Shibli, M. M. A. Pritom, and M. Gupta. Abusegpt: Abuse of generative ai chatbots to create smishing campaigns. In *Int. Symp. on Dig. Forensics and Security (ISDFS)*, pages 1–6, 2024.

[25] B. Yang, G. Chen, and X. Luo. A legal judgment prediction model based on bert, attention, and graph convolutional network. In *2024 IEEE 36th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 303–310, 2024.

[26] J. Yoo and Y. Cho. Icsa: Intelligent chatbot security assistant using text-cnn and multi-phase real-time defense against sns phishing attacks. *Exp. Syst. with Applications*, 207:117893, 2022.

[27] L. Zheng, N. Guha, B. R. Anderson, P. Henderson, and D. E. Ho. When does pretraining help? assessing self-supervised learning for law and the casehold dataset of 53,000+ legal holdings. In *Int. conf. on AI and law*, pages 159–168, 2021.