КІБЕРБЕЗПЕКА: освіта, наука, техніка
CYBERSECURITY: EDUCATION, SCIENCE, TECHNIQUE
№ 4 (32), 2026
ISSN 2663 – 4023

DOI 10.28925/2663-4023.2026.32.1196
УДК 004.056.5

Taras Fedynyshyn
Ph.D. student in the Department of Information Protection
Lviv Polytechnic National University, Lviv, Ukraine
ORCID: 0009-0006-8233-8057
fedynyshyn.taras@gmail.com

Olha Partyka
Ph.D. in Physics and Mathematics, Associate Professor in the Department of Information Protection
Lviv Polytechnic National University, Lviv, Ukraine
ORCID: 0000-0002-3086-3160
olha.o.mykhailova@lpnu.ua

RETRIEVAL-AUGMENTED GENERATION FOR FORENSIC LEGAL ANALYSIS: INTEGRATION OF UKRAINIAN CRIMINAL CODE WITH MOBILE DEVICE EVIDENCE

Abstract. Digital forensic investigations in Ukraine require analysts to classify mobile device evidence according to the Criminal Code, a process that is time-consuming and requires deep legal expertise. This paper presents the first retrieval-augmented generation (RAG) system for Ukrainian Criminal Code analysis, focusing on Section I (Crimes Against National Security). We construct a database of 9 articles covering treason, espionage, collaboration, and sabotage offenses, and evaluate the system on 60 synthetic forensic scenarios with deterministically-derived ground truth. Our experiments compare four chunking strategies, three multilingual embedding models, and four large language models (both API-based and locally-deployed). The best retrieval configuration achieves MRR of 0.588 using multilingual-e5-large embeddings with part-level chunking. For end-to-end classification, RAG with GPT-4o-mini achieves 54.2% article identification accuracy, outperforming a few-shot prompting baseline (29.2%, p=0.03) but showing no statistically significant improvement over direct LLM prompting (52.1%, p=0.89). We argue that RAG's primary advantage for forensic applications lies not in classification accuracy but in grounding, transparency, and governance: retrieved legal provisions are traceable and verifiable, the knowledge base can be updated without retraining, and the system supports fully local deployment where evidence cannot leave the organization. Local LLMs achieve 77% of API performance (41.7% accuracy), confirming that on-premise deployment is feasible at reduced accuracy.

Keywords: Retrieval-Augmented Generation; Legal NLP; Ukrainian Criminal Code; Digital Forensics; Multilingual Embeddings; SLM; LLM, Mobile Forensics.

INTRODUCTION

Mobile devices seized during criminal investigations contain evidence that must be classified according to applicable criminal law. In Ukraine, forensic analysts manually match digital artifacts-messages, call logs, location data, images-against 450+ Criminal Code articles [1]. This process is time-consuming and requires deep legal expertise. Recent work has explored AI-driven approaches to mobile forensics [2], but mapping forensic artifacts to specific legal provisions remains largely manual. For national security offenses, where Section I defines crimes such as treason, espionage, and sabotage, accurate classification is both legally and politically sensitive.

Large language models (LLMs) have shown strong performance on legal reasoning tasks in English [3], but their application to Ukrainian legal text remains unexplored.

Ukrainian is a low-resource language for NLP, and the Criminal Code uses domain-specific terminology that general-purpose models handle poorly. Fine-tuning LLMs on Ukrainian legal data requires substantial annotated corpora that do not exist. Retrieval-Augmented Generation (RAG) [4] offers an alternative: instead of encoding legal knowledge in model parameters, the system retrieves relevant law articles at inference time and provides them as context. Beyond potential accuracy improvements, RAG has properties particularly relevant in governance-sensitive forensic environments: classifications are grounded in specific retrieved provisions rather than opaque model internals, the legal knowledge base can be updated when legislation changes without retraining, and the full pipeline can run locally on consumer hardware-a requirement when classified evidence must not leave the organization.

Article objective: The objective of this paper is to develop and evaluate a retrieval-augmented generation (RAG) system for automated classification of mobile forensic evidence according to the Ukrainian Criminal Code, enabling efficient and transparent legal analysis in digital forensic investigations. Achieving this objective requires solving the following tasks: (1) construct a structured database of Ukrainian Criminal Code Section I (Crimes Against National Security) with bilingual text, metadata, and cross-references suitable for vector retrieval; (2) evaluate chunking strategies (article-level, part-level, semantic) for Ukrainian legal text to determine optimal granularity for retrieval; (3) compare multilingual embedding models (E5-large, E5-base, MPNet) on Ukrainian legal text retrieval performance; (4) assess end-to-end RAG classification accuracy using both cloud-based (GPT-4o-mini) and locally-deployed (Llama-3, Qwen2.5, Gemma-3) language models; (5) compare RAG against baseline approaches (direct LLM prompting, few-shot prompting) to quantify retrieval's contribution to classification accuracy; (6) analyze error patterns to identify limitations and inform future improvements.

The contributions of this paper are: (1) A RAG system architecture for Ukrainian Criminal Code retrieval, including a structured database of Section I articles with metadata and cross-references. (2) An evaluation of embedding models and chunking strategies for Ukrainian legal text, showing multilingual embeddings with part-level chunking outperform article-level and sentence-level approaches. (3) A comparison of RAG, direct LLM prompting, and few-shot prompting for legal classification, showing RAG outperforms few-shot prompting (54.2% vs. 29.2%, p=0.03) but not direct prompting (52.1%). (4) An analysis of local versus API-based LLMs for forensic applications, with local models achieving 41.7% accuracy-showing feasibility as decision-support where data cannot leave the organization.

Analysis of recent research and publications. RAG combines parametric knowledge in language model weights with non-parametric knowledge retrieved from external corpora [4], addressing hallucination by grounding generation in retrieved evidence. Retrieval has evolved from sparse methods like BM25 [5] to dense approaches. Dense Passage Retrieval [6] established that learned embeddings outperform sparse methods by 9-19% on open-domain QA. Sentence-BERT [7] made practical sentence-level retrieval feasible.

Legal text poses unique NLP challenges: specialized terminology, complex syntax, and heavy cross-references [8]. Domain-specific pretraining improves performance-Legal-BERT [9] gained 2-5% F1 over general BERT on legal classification. Recent legal RAG work reveals domain-specific difficulties. Reuter et al. [10] found retrieval precision degrades with corpus size due to high similarity between legal provisions. Ho et al. [11] showed encoding hierarchical legal structure into RAG improves reasoning for multi-factor legal tests.

Multilingual embedding models represent text from different languages in a shared semantic space, important for low-resource languages with limited training data. LaBSE [12] was trained on parallel data from 109 languages using translation ranking. Wang et al. [13] introduced Multilingual E5, trained with contrastive learning on large multilingual corpora.

Digital forensics involves identification, preservation, analysis, and presentation of digital evidence. Lillis et al. [14] identified key challenges: increasing data volume, encryption, cloud storage, and need for faster analysis. Dunsin et al. [15] surveyed AI and machine learning in digital forensics and incident response. Mykhaylova et al. [2] proposed an AI-driven roadmap for person-of-interest detection on mobile forensics data, demonstrating how machine learning assists in identifying relevant evidence patterns.

Two main approaches adapt language models to specialized domains: retrieval-augmented generation and fine-tuning. Hu et al. [16] introduced LoRA, reducing fine-tuning computational cost by learning low-rank weight matrix updates.

Despite progress in legal NLP and RAG systems, several gaps remain. First, no prior work has developed RAG systems for Ukrainian legal text. Existing legal RAG research focuses on English, German, and Chinese legal systems, leaving Slavic languages unexplored. Second, the intersection of digital forensics and legal RAG has not been studied. Third, multilingual embedding model performance on Ukrainian legal text is unknown.

THEORETICAL FOUNDATIONS

Retrieval-augmented generation. RAG combines parametric knowledge in language model weights with non-parametric knowledge retrieved from external corpora [4]. The architecture consists of two components: a retriever that identifies relevant documents given a query, and a generator that produces output conditioned on both the query and retrieved documents. This approach addresses LLM hallucination by grounding generation in retrieved evidence.

Dense retrieval methods encode queries and documents as vectors in a shared embedding space, enabling efficient similarity search. Dense Passage Retrieval [6] demonstrated that learned embeddings outperform sparse methods by 9–19% on open-domain QA tasks. Sentence-BERT [7] made practical sentence-level retrieval feasible by producing fixed-length sentence embeddings.

Legal NLP challenges. Legal text poses unique NLP challenges that distinguish it from general-domain text [8]. Legal language exhibits specialized terminology, complex syntactic structures, and extensive cross-references between provisions. Domain-specific pretraining improves performance-Legal-BERT [9] achieved 2-5% F1 improvement over general BERT on legal classification tasks.

For Ukrainian legal text, additional challenges arise. Ukrainian is a morphologically complex language with limited NLP resources. The Criminal Code uses domain-specific terminology that differs from everyday language, and legal texts contain terms borrowed from Russian legal tradition or translated from European frameworks. No embedding models have been specifically trained or evaluated on Ukrainian legal text.

Legal NLP challenges. Multilingual embedding models represent text from different languages in a shared semantic space. LaBSE [12] was trained on parallel data from 109 languages using translation ranking, producing embeddings where semantically similar sentences cluster regardless of language. Multilingual E5 [13] was trained with contrastive learning on large multilingual corpora, achieving strong cross-lingual retrieval performance across 100+ languages. For Slavic languages, most multilingual models allocate limited capacity compared to English and Western European languages. Ukrainian, despite 40+ million speakers, is underrepresented in model training data, creating a gap between theoretical multilingual capabilities and actual Ukrainian text performance.

RESEARCH METHODOLOGY

Criminal code database construction. We constructed a structured database from Section I of the Ukrainian Criminal Code (Crimes Against National Security), obtained from zakon.rada.gov.ua [1]. This section was selected for its relevance to Security Service of Ukraine investigations, clear statutory definitions enabling deterministic ground truth, and inclusion of 2022 wartime amendments.

The database covers 9 articles with 25 constituent parts: Article 109 (Forceful change of constitutional order, 3 parts), Article 110 (Encroachment on territorial integrity, 3 parts), Article 111 (High treason, 2 parts), Article 111-1 (Collaborative activity, 7 parts-added 2022), Article 111-2 (Aiding the aggressor state, 3 parts-added 2022), Article 112 (Assassination of state officials, 1 part), Article 113 (Sabotage, 2 parts), Article 114 (Espionage, 2 parts), and Article 114-1 (Obstruction of Armed Forces activities, 2 parts).

We evaluated four chunking strategies: article-level chunking (9 chunks), part-level chunking (25 chunks), and two semantic chunking variants (512-token and 256-token maximum with overlap).

Forensic scenario dataset. We created UCC-Forensic-60, a dataset of 60 synthetic forensic scenarios representing mobile device evidence from national security investigations, split into 12 development and 48 test scenarios. Scenarios are distributed across four categories: TREASON_ESPIONAGE (Articles 111, 114), COLLABORATION (Articles 111-1, 111-2), SUBVERSION (Articles 109, 110, 113), and TARGETED_VIOLENCE (Articles 112, 114-1).

Ground truth labels were derived deterministically from statutory language. Each scenario contains specific factual elements mapping directly to article definitions, eliminating the need for expert annotators.

RAG system architecture. Our RAG pipeline follows the standard retrieve-then-generate approach [4]. Given a forensic scenario description as input, the system retrieves relevant Criminal Code chunks and generates a legal classification with the retrieved context.

We evaluated three multilingual embedding models: Multilingual-E5-large [13] (1024-dimensional), Multilingual-E5-base (768-dimensional), and Paraphrase-multilingual-mpnet [7] (768-dimensional). We use FAISS [17] for vector storage with IndexFlatIP. We retrieve top-k chunks (k in {1,3,5}) ranked by cosine similarity.

We evaluated four language models: GPT-4o-mini (OpenAI API), Llama-3-8B-Instruct, Qwen2.5-Coder-7B-Instruct, and Gemma-3-12B-Instruct (local via llama.cpp with Q4_K_M quantization on Apple M3 Max).

Evaluation Framework. We evaluate retrieval quality using standard information retrieval metrics [18]. Mean Reciprocal Rank (MRR) is the average of reciprocal ranks of the first relevant chunk. Precision@k is the fraction of retrieved chunks (among top-k) that are relevant. Recall@k is the fraction of relevant chunks that appear in top-k results.

For the generation component, Article Accuracy is the fraction of exact matches with ground truth. Multi-label F1 is micro-averaged F1 for multi-article scenarios. All metrics include 95% confidence intervals via bootstrap resampling (1,000 iterations).

RESEARCH RESULTS

Chunking strategy comparison. Table 1 presents retrieval performance across four chunking strategies.

Table 1

Retrieval performance by chunking strategy

Strategy	Chunks	MRR	P@1	P@3
part_level	25	0.524	0.438	0.583
article_level	9	0.470	0.292	0.531
semantic_512	10	0.434	0.250	0.531
semantic_256	11	0.418	0.271	0.531

Part-level chunking achieved the highest MRR (0.524), outperforming article-level by 0.054 points. The difference was not statistically significant (p=0.25, d=0.13), but the trend was consistent. Semantic chunking underperformed structure-aware approaches, demonstrating that respecting legal document structure provides better retrieval than arbitrary token-based splitting.

Embedding model comparison. Table 2 compares three multilingual embedding models using part-level chunking.

Table 2

Retrieval performance by embedding model

Model	Dim	MRR	P@1	P@3
multilingual-e5-large	1024	0.588	0.542	0.563
multilingual-e5-base	768	0.524	0.438	0.583
paraphrase-multilingual-mpnet	768	0.439	0.333	0.490

Multilingual-e5-large achieved the highest MRR (0.588) and Precision@1 (0.542). The best configuration (part-level chunking with multilingual-e5-large) achieves MRR of 0.588, meaning the relevant article typically appears in the top two results-acceptable for analyst decision support workflows.

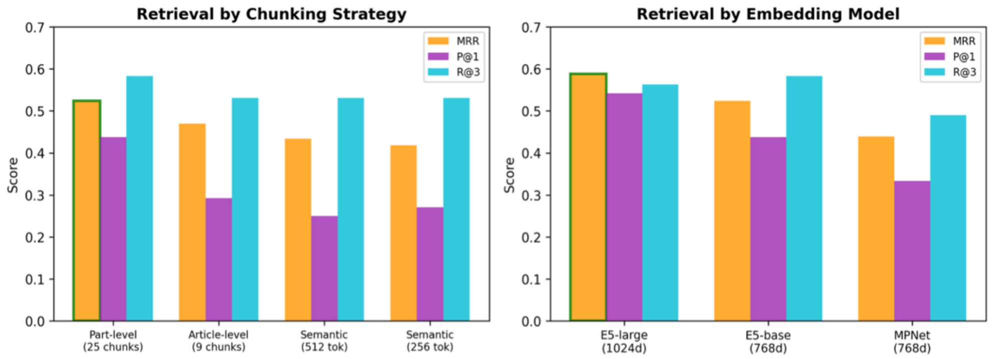

Figure 1. Retrieval performance comparison across embedding models and chunking strategies. Left: MRR by chunking strategy (part-level outperforms semantic chunking). Right: MRR by embedding model (multilingual-e5-large achieves best performance at 0.588)

End-to-End Generation Results. Table 3 presents end-to-end performance using the best retrieval configuration with four generation models.

Table 3

End-to-end RAG performance by generation model

Model	Type	Article Acc	ML-F1	Parse	Latency
gpt-4o-mini	API	54.2%	46.9%	100%	651ms
qwen2.5-coder-7b	local	41.7%	35.4%	93.8%	1901ms
llama3-8b	local	39.6%	38.2%	89.6%	1286ms
gemma3-12b	local	39.6%	28.5%	100%	1840ms

GPT-4o-mini achieved the highest accuracy (54.2%) and multi-label F1 (46.9%), with perfect parse success and lowest latency. Among local models, Qwen2.5-coder-7b achieved highest accuracy (41.7%). The 12.5 percentage point gap between cloud and local models reflects both model capability and the challenge of running quantized 7–8B models on legal classification in low-resource languages.

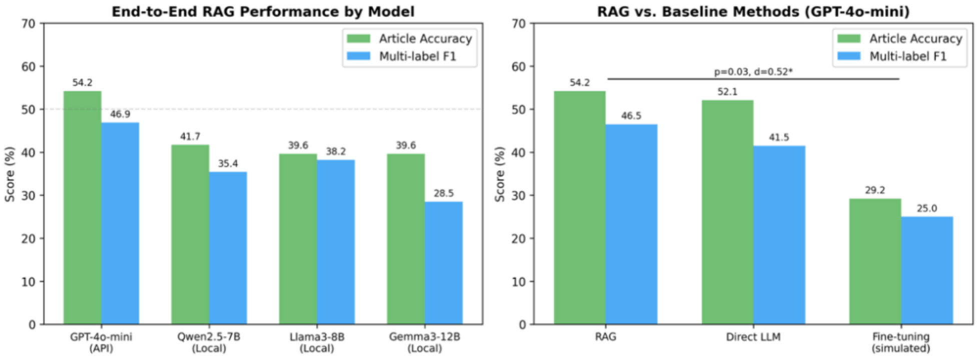

Figure 2. Experimental results summary. Left: End-to-end RAG performance across generation models, showing the accuracy gap between cloud API (GPT-4o-mini) and local models. Right: Comparison of RAG against baseline methods, with significant improvement over the few-shot prompting baseline (p=0.03, Cohen's d=0.52).

RAG vs. baseline comparison. Table 4 compares our RAG system against two baselines.

Table 4

Comparison of RAG against baseline approaches

Method	Article Acc	ML-F1	Latency
RAG	54.2%	46.5%	713ms
Direct LLM	52.1%	41.5%	594ms
Few shot prompting	29.2%	25.0%	506ms

RAG achieved 54.2% accuracy versus 52.1% for direct prompting-a 2.1 percentage point difference that was not statistically significant (p=0.89, d=0.04). However, RAG substantially outperformed few-shot prompting: 54.2% vs. 29.2%-a 25.0 percentage point difference that was statistically significant (p=0.03, Cohen's d=0.52, medium effect).

Error Analysis. Analysis of all 22 incorrect GPT-4o-mini predictions revealed that reasoning errors (59.1%, 13 cases) were more common than retrieval errors (40.9%, 9 cases).

Most frequent reasoning errors involved confusion between semantically similar articles: Article 111 vs. 111-1 (5 cases), Article 111-1 vs. 111-2 (4 cases), and Article 114 vs. 111 (3 cases). These patterns reflect genuine legal complexity-2022 amendments created overlapping provisions that even experts may interpret differently.

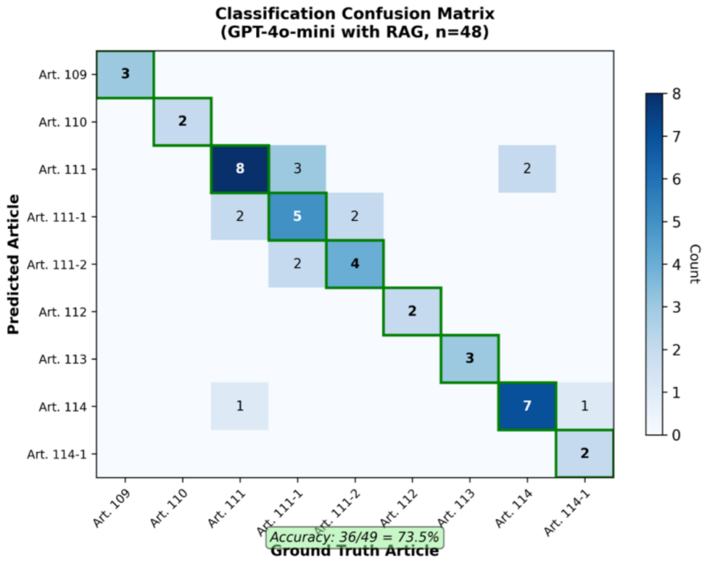

Figure 3. Classification confusion matrix for end-to-end RAG with GPT-4o-mini on the 48-scenario test set. Rows represent ground truth articles; columns represent predicted articles. Diagonal cells indicate correct classifications. The matrix shows systematic confusion between related articles, particularly within the collaboration cluster (Articles 111, 111-1, 111-2)

Summary of findings. Our experiments yield four main findings: (1) Part-level chunking outperforms semantic chunking for Ukrainian Criminal Code retrieval by respecting legal document structure. (2) Multilingual-e5-large achieves best retrieval (MRR 0.588). (3) Cloud LLMs outperform local models (GPT-4o-mini: 54.2% vs. local: 39.6–41.7%). (4) RAG outperforms few-shot prompting (54.2% vs. 29.2%, p=0.03, d=0.52) but not direct LLM prompting (52.1%, p=0.89), suggesting RAG's value lies in grounding, transparency, and updatability rather than accuracy gains for small corpora.

CONCLUSIONS AND PROSPECTS FOR FURTHER RESEARCH

This paper presented a RAG system for Ukrainian Criminal Code analysis in digital forensics contexts. We focused on Section I (Crimes Against National Security), developing both retrieval infrastructure and an evaluation framework using statutory language mapping as ground truth.

Main findings. Multilingual-e5-large with part-level chunking achieved highest retrieval performance (MRR = 0.588). Part-level chunking worked better than article-level or semantic approaches by preserving legal semantics of individual offense definitions. RAG with GPT-4o-mini achieved 54.2% accuracy, outperforming few-shot prompting (29.2%, p=0.03) but showing no significant improvement over direct LLM prompting (52.1%, p=0.89). Local models on consumer hardware achieved approximately 77% of API performance (41.7%), representing a concrete privacy-accuracy trade-off. The lack of significant accuracy difference between RAG and direct prompting does not undermine retrieval's value. RAG provides three independent advantages: (1) grounding and transparency-analysts can verify the legal basis for classifications; (2) updatability-the knowledge base can incorporate legislative changes without retraining; (3) local deployment-the entire pipeline can run on-premise where classified evidence must remain within the organization.

Limitations. All scenarios were synthetically generated with deterministic ground truth, making classification artificially well-formed compared to real forensic evidence. We evaluated only Section I (9 articles); generalization requires validation on other sections. The 48-scenario sample provides limited statistical power.

Prospects for further research. Several directions could extend this work: (1) Evaluating on the full Criminal Code (450+ articles) to test whether retrieval degrades with corpus size. (2) Partnership with forensic practitioners to obtain anonymized real evidence. (3) Combining RAG with lightweight fine-tuning on retrieved context. (4) Training embedding models specifically for Ukrainian legal text. (5) Developing confidence-based abstention mechanisms for low-confidence predictions. (6) Creating severity-weighted error metrics based on penalty distance between misclassified articles. The broader implication is that RAG provides a viable architectural framework for applying LLMs to low-resource legal domains. The governance advantages-grounding in traceable legal text, updatability when legislation changes, and support for on-premise deployment-are especially relevant for forensic applications where classification decisions must be auditable and evidence must remain within organizational boundaries.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Verkhovna Rada of Ukraine. (2001). *Criminal Code of Ukraine.* https://zakon.rada.gov.ua/laws/show/2341-14
2. Mykhaylova, O., Fedynyshyn, T., Sokolov, V., & Kyrychok, R. (2024). Person-of-interest detection on mobile forensics data: AI-driven roadmap. *CEUR Workshop Proceedings*, *3654*, 239–252. https://ceur-ws.org/Vol-3654/paper20.pdf
3. Brown, T., et. al., (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901. https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html
4. Lewis, P., et. al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, *33*, 9459-9474. https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html
5. Robertson, S., & Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, *3*(4), 333–389. https://doi.org/10.1561/1500000019
6. Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., … Yih, W. (2020). Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* (pp. 6769–6781). https://doi.org/10.18653/v1/2020.emnlp-main.550
7. Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing* (pp. 3982–3992). https://doi.org/10.18653/v1/D19-1410
8. Zhong, H., Xiao, C., Tu, C., Zhang, T., Liu, Z., & Sun, M. (2020). How does NLP benefit legal systems: A summary of legal artificial intelligence. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5218–5230). https://doi.org/10.18653/v1/2020.acl-main.466
9. Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., & Androutsopoulos, I. (2020). LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 2898–2904). https://doi.org/10.18653/v1/2020.findings-emnlp.261
10. Reuter, M., Lingenberg, T., Liepina, R., Lagioia, F., Lippi, M., Sartor, G., Passerini, A., & Sayin, B. (2025). Towards reliable retrieval in RAG systems for large legal datasets. In *Proceedings of the Natural Legal Language Processing Workshop 2025*. https://doi.org/10.18653/v1/2025.nllp-1.3

11. Ho, J., Colby, A., & Fisher, W. (2025). Incorporating legal structure in retrieval-augmented generation: A case study on copyright fair use. *arXiv*. https://doi.org/10.48550/arXiv.2505.02164
12. Feng, F., Yang, Y., Cer, D., Arivazhagan, N., & Wang, W. (2022). Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics* (pp. 878–891). https://doi.org/10.18653/v1/2022.acl-long.62
13. Wang, L., Yang, N., Huang, X., Yang, L., Majumder, R., & Wei, F. (2024). Multilingual E5 text embeddings: A technical report. *arXiv*. https://doi.org/10.48550/arXiv.2402.05672
14. Lillis, D., Becker, B., O'Sullivan, T., & Scanlon, M. (2016). Current challenges and future research areas for digital forensic investigation. In *Proceedings of the Annual ADFSL Conference on Digital Forensics, Security and Law* (pp. 9–20). https://commons.erau.edu/adfsl/2016/tuesday/5/
15. Dunsin, D., Ghanem, M. C., Ouazzane, K., & Vassilev, V. (2023). Artificial intelligence and machine learning in digital forensics and incident response. *Forensic Science International: Digital Investigation, 48*, 301675. https://doi.org/10.1016/j.fsidi.2023.301675
16. Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2022). LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR 2022)*. https://openreview.net/forum?id=nZeVKeeFYf9
17. Johnson, J., Douze, M., & Jégou, H. (2019). Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, *7*(3), 535–547. https://doi.org/10.1109/TBDATA.2019.2921572
18. Voorhees, E. M. (1999). The TREC-8 question answering track report. In *Proceedings of the Eighth Text REtrieval Conference (TREC-8)* (pp. 77–82). https://trec.nist.gov/pubs/trec8/papers/qa_report.pdf

Фединишин Тарас Олегович
аспірант кафедри захисту інформації
Національний Університет «Львівська Політехніка», Львів, Україна
ORCID: 0009-0006-8233-8057
fedynyshyn.taras@gmail.com

Партика Ольга Олександрівна
к.ф-м.н., доцент кафедри захисту інформації
Національний університет «Львівська Політехніка», Львів, Україна
ORCID: 0000-0002-3086-3160
olha.o.mykhailova@lpnu.ua

ГЕНЕРАЦІЯ З ДОПОВНЕННЯМ НА ОСНОВІ ПОШУКУ ДЛЯ СУДОВО-ПРАВОВОГО АНАЛІЗУ: ІНТЕГРАЦІЯ КРИМІНАЛЬНОГО КОДЕКСУ УКРАЇНИ З ДОКАЗАМИ З МОБІЛЬНИХ ПРИСТРОЇВ

Анотація. Цифрові криміналістичні розслідування в Україні вимагають від аналітиків класифікації доказів, отриманих із мобільних пристроїв, відповідно до положень Кримінального кодексу, що є трудомістким процесом і потребує глибокої правової експертизи. У цій статті представлено першу систему генерації з доповненням вибіркою (retrieval-augmented generation, RAG) для аналізу Кримінального кодексу України з акцентом на Розділ I («Злочини проти основ національної безпеки України»). Було сформовано базу даних із 9 статей, що охоплюють державну зраду, шпигунство, колабораційну діяльність і диверсію, та здійснено оцінювання системи на основі 60 синтетичних криміналістичних сценаріїв із детерміновано сформованою еталонною розміткою (ground truth). У межах експериментів порівнювалися чотири стратегії сегментації тексту, три багатомовні моделі ембедингів і чотири великі мовні моделі (як API-орієнтовані, так і розгорнуті локально). Найкраща конфігурація пошуку досягла показника MRR на рівні 0,588 із використанням ембедингів multilingual-e5-large та сегментації на рівні частин статей. Для наскрізної класифікації RAG у поєднанні з GPT-4o-mini забезпечила точність ідентифікації статті на рівні 54,2 %, перевищивши базову модель few-shot prompting (29,2 %, p = 0,03), проте не продемонструвала статистично значущого покращення порівняно з прямим застосуванням LLM (52,1 %, p = 0,89). Обґрунтовується, що основна перевага RAG для криміналістичних застосувань полягає не стільки в підвищенні точності класифікації, скільки у забезпеченні обґрунтованості, прозорості та належного управління: витягнуті правові норми є відстежуваними та верифікованими, база знань може оновлюватися без перенавчання моделі, а система підтримує повністю локальне розгортання у випадках, коли докази не можуть залишати межі організації. Локальні великі мовні моделі досягають 77 % продуктивності API-рішень (41,7 % точності), що підтверджує можливість локального розгортання за умови певного зниження точності.

Ключові слова: Retrieval-Augmented Generation; Legal NLP; Ukrainian Criminal Code; Digital Forensics; Multilingual Embeddings; SLM; LLM, Mobile Forensics.

REFERENCES (TRANSLATED AND TRANSLITERATED)

1. Verkhovna Rada of Ukraine. (2001). *Criminal Code of Ukraine.* https://zakon.rada.gov.ua/laws/show/2341-14
2. Mykhaylova, O., Fedynyshyn, T., Sokolov, V., & Kyrychok, R. (2024). Person-of-interest detection on mobile forensics data: AI-driven roadmap. *CEUR Workshop Proceedings, 3654,* 239–252. https://ceur-ws.org/Vol-3654/paper20.pdf
3. Brown, T., et. al., (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33,* 1877–1901. https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

4. Lewis, P., et. al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems, 33*, 9459-9474. https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html
5. Robertson, S., & Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval, 3*(4), 333–389. https://doi.org/10.1561/1500000019
6. Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., … Yih, W. (2020). Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* (pp. 6769–6781). https://doi.org/10.18653/v1/2020.emnlp-main.550
7. Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing* (pp. 3982–3992). https://doi.org/10.18653/v1/D19-1410
8. Zhong, H., Xiao, C., Tu, C., Zhang, T., Liu, Z., & Sun, M. (2020). How does NLP benefit legal systems: A summary of legal artificial intelligence. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5218–5230). https://doi.org/10.18653/v1/2020.acl-main.466
9. Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., & Androutsopoulos, I. (2020). LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 2898–2904). https://doi.org/10.18653/v1/2020.findings-emnlp.261
10. Reuter, M., Lingenberg, T., Liepina, R., Lagioia, F., Lippi, M., Sartor, G., Passerini, A., & Sayin, B. (2025). Towards reliable retrieval in RAG systems for large legal datasets. In *Proceedings of the Natural Legal Language Processing Workshop 2025*. https://doi.org/10.18653/v1/2025.nllp-1.3
11. Ho, J., Colby, A., & Fisher, W. (2025). Incorporating legal structure in retrieval-augmented generation: A case study on copyright fair use. *arXiv*. https://doi.org/10.48550/arXiv.2505.02164
12. Feng, F., Yang, Y., Cer, D., Arivazhagan, N., & Wang, W. (2022). Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics* (pp. 878–891). https://doi.org/10.18653/v1/2022.acl-long.62
13. Wang, L., Yang, N., Huang, X., Yang, L., Majumder, R., & Wei, F. (2024). Multilingual E5 text embeddings: A technical report. *arXiv*. https://doi.org/10.48550/arXiv.2402.05672
14. Lillis, D., Becker, B., O’Sullivan, T., & Scanlon, M. (2016). Current challenges and future research areas for digital forensic investigation. In *Proceedings of the Annual ADFSL Conference on Digital Forensics, Security and Law* (pp. 9–20). https://commons.erau.edu/adfsl/2016/tuesday/5/
15. Dunsin, D., Ghanem, M. C., Ouazzane, K., & Vassilev, V. (2023). Artificial intelligence and machine learning in digital forensics and incident response. *Forensic Science International: Digital Investigation, 48*, 301675. https://doi.org/10.1016/j.fsidi.2023.301675
16. Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2022). LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR 2022)*. https://openreview.net/forum?id=nZeVKeeFYf9
17. Johnson, J., Douze, M., & Jégou, H. (2019). Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data, 7*(3), 535–547. https://doi.org/10.1109/TBDATA.2019.2921572
18. Voorhees, E. M. (1999). The TREC-8 question answering track report. In *Proceedings of the Eighth Text REtrieval Conference (TREC-8)* (pp. 77–82). https://trec.nist.gov/pubs/trec8/papers/qa_report.pdf

Отримано редакцією журналу / Received: 16.01.26
Прорецензовано / Revised: 02.02.26
Схвалено до друку / Accepted: 26.03.26