Addressing concept drift in 5G CVE classification with LLMs

Pierpaolo Bene[1], Andrea Bernardini[2,*,†], Leonardo Sagratella[2,†] and Nicolò Maunero[3]

[1]Department of Control and Computer Engineering, Politecnico di Torino, Turin, Italy
[2]Fondazione Ugo Bordoni, Viale del Policlinico, 147, 00161, Rome, Italy
[3]IMT School for Advanced Studies, Lucca, Italy

Abstract

CVE disclosures are projected to reach 50,000 in 2025, and with NIST's National Vulnerability Database facing ongoing operational issues, organizations struggle to identify which vulnerabilities pose actual risks to their systems. This creates an urgent need for automated systems, particularly in rapidly evolving domains such as 5G, where new attack surfaces and vulnerability patterns emerge frequently. A key obstacle is concept drift, the evolution of CVE descriptions and technical terminology over time. Static classification models trained on historical data degrade in performance when applied to newly published vulnerabilities, as the underlying data distribution shifts. In this paper, we evaluate four classification strategies for identifying 5G related CVEs: embedding-based classification with an SVM, large-scale LLMs as Qwen3-32B, quantized compact models Qwen3-4B (4-bit), and a distillation-based compact model (Qwen3-4B with distilled reasoning chains from 32B using *LoRA*). Models are trained on a newly published dataset of 510 manually labeled 5G CVEs by experts and tested on an out-of-distribution set of 123 new published CVEs.

While keyword-based filtering constrains the dataset to explicitly labeled vulnerabilities, it enables controlled evaluation of models' performance on well-defined domain-specific instances.

The results indicate that, for specialized security classification of 5G vulnerabilities, compact quantized models provide practical advantages in computational efficiency and resource utilization without compromising precision. The Qwen3-4B (4-bit) achieves performance comparable to the much larger Qwen3-32B model and avoids the hallucination issues observed in distilled model variants. In contrast, SVMs exhibit significant performance degradation when applied to out-of-distribution data.

A key advantage of LLM-based approaches is their interpretability. By generating structured reasoning chains, these models provide transparent decision-making processes that allow security analysts to inspect and validate the motivations behind each classification.

Keywords

5G, CVE, Vulnerabilities, LLM, SVM, Security, Low-Rank Adaptation, LoRA, Classification, Embeddings,

1. Introduction

The National Vulnerability Database (NVD) has experienced repeated funding disruptions in 2025, resulting in a growing backlog of CVEs to analyze.

In 2025, at least 50,000 new Common Vulnerabilities and Exposures (CVEs) were expected, and more than 26,000 are already waiting for analysis [1] [2]. In addition to that, around 40% of reported vulnerabilities lack complete Common Vulnerability Scoring System (CVSS) assessments or have incomplete initial reports [3]. Clearly, this backlog is a growing problem for cybersecurity, as the number and severity of vulnerabilities increase. Moreover, about 38% of CVEs reported in the first half of 2025 were rated High or Critical with a CVSS score above 7.0, according to [4].

On the other hand, traditional methods for managing and classifying vulnerabilities, such as keyword filtering and manual review, are slow, error-prone, and cannot handle the continuous publication of new vulnerabilities along with the evolution of technologies.

Joint National Conference on Cybersecurity (ITASEC SERICS 2026), February 09-13, 2026, Cagliari, IT
*Corresponding author.
†These authors contributed equally.
pierpaolo.bene@studenti.polito.it (P. Bene); abernardini@fub.it (A. Bernardini); lsagratella@fub.it (L. Sagratella); nicolo.maunero@imtlucca.it (N. Maunero)
© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-04-23

CEUR-WS.org/Vol-4198/paper16.pdf

These challenges are especially pronounced in the context of 5G networks, an ecosystem that has evolved significantly between 2019-2024, as demonstrated by this timeline:

- 2019–2020: Initial deployment of foundational 5G components, including the gNodeB (gNB), 5G New Radio (NR), and early implementations of the 5G Core (both Non-Standalone and Standalone configurations), following the specifications of 3GPP Releases 15 and 16.
- 2021–2022: Consolidation of network virtualization strategies, marked by a transition from traditional NFV/VM-based network functions to cloud-native architectures. During this phase, containerization and the adoption of Kubernetes-based orchestration accelerated the shift toward Cloud-Native Network Functions (CNFs) in commercial 5G deployments.
- 2023–2024: Emergence and maturation of advanced technologies such as Open Radio Access Networks (Open RAN), the introduction of 5G-Advanced as defined in 3GPP Release 18, and the integration of AI/ML techniques for network automation, energy efficiency, anomaly detection, and intelligent RAN control.

For example, a CVE affecting Kubernetes containers in 2019 would typically not be classified as 5G-relevant at that time, when 5G networks primarily relied on dedicated network appliances rather than containerized deployments.

However, by 2023, since cloud-native 5G architectures became the standard, such CVEs can be identified as relevant to 5G. This suggests that the relationship between CVE descriptions (*features*) and their 5G-relevance classification (*labels*) changes over time, the so-called *concept drift*. In terms of distributions, it can be expressed as $P_{2019}(Y|X) \neq P_{2023}(Y|X)$, meaning that the underlying relationship between features and labels has changed over time.

While concept drift [5, 6] is a well-documented phenomenon in healthcare monitoring [7], credit card fraud detection [8], and manufacturing process control [9], its manifestation in cybersecurity, specifically within CVE terminology evolution, remains relatively underexplored [10].

2. Related works

Methodologies for classifying vulnerability descriptions have gone through several stages of development. Initial approaches were built on traditional feature engineering, extracting representations like term TF-IDF. These features were then used as inputs to standard machine learning classifiers, such as Support Vector Machines (SVMs) and Random Forests.

Subsequently, a further approach employed pre-trained language models (BERT, RoBERTa) for feature extractors, and trained lightweight classifiers on top of these contextual embeddings. A refinement came with sentence transformers such as all-MiniLM-L12-v2 [11], which are BERT-based models fine-tuned specifically to produce high-quality sentence embeddings. These models simplify deployment while achieving strong performance when paired with simple classifiers [12, 13]. However, small LLMs (<10B parameters) consistently outperform previous approaches in classification tasks [14] even by improving the context of LLMs as in [15] and in [16].

Regarding CVEs, Ghosh et al. [17] fine-tuned open-source LLMs for vulnerability assessment of medical devices, demonstrating that domain-adapted models can effectively perform this specialized task. Authors in [18] investigate the effectiveness of Large Language Models in generating CVSS vectors from CVE descriptions. Their work compares LLM-based approaches against embedding-based methods using supervised learning classifiers, and proposes a hybrid approach that combines LLMs for objective components and embedding-based classifiers for subjective components. A systematic review of the use of LLMs in cybersecurity is presented in [19], by analyzing more than 300 works exploring 10 different scenarios, among which vulnerability detection is directly related to the problem of CVE classification. This work highlights the high potential of LLMs for this task while also showing their limitations, particularly the lack of datasets for specific domains.

However, existing embedding-based and fine-tuning approaches typically assume static data distributions. This assumption breaks down when CVE descriptions exhibit concept drift [10], where new attack vectors introduce terms that were absent during classifier training.

To address this challenge, recent work shows that distilling reasoning processes can improve model robustness to distribution shifts [20, 21]

2.1. Our Contributions

This paper contributes to 5G vulnerability classification under concept drift through:

- **Embeddings based Support Vector Machine (SVM)**: For the classification baseline, we extract semantic embeddings using *all-MiniLM-L12-v2* [11] and evaluate a trained SVM using an RBF kernel as a reference for comparison with LLM-based models.
- **Training Data Augmentation and Fine-Tuning Strategy**: We distill classification capabilities and reasoning from *Qwen3-32B* to a compact 4-bit quantized model *Qwen3-4B(4 bit)* by generating synthetic reasoning that explains classification decisions on 5G related CVE, then fine-tuned the smaller model on this augmented dataset using parameter-efficient *Low-Rank Adaptation* (LoRA)[22].
- **Evaluation on new published CVEs**: We conduct a preliminary comparison of embedding-based SVM, base quantized models, and various LoRA fine-tuned models on recent CVEs (published between July 1st and September 13th, 2025). Initial results show that the *Qwen3-4B(4 bit)* model achieves a high precision on 123 newly published CVEs, while embedding-based SVM classifiers exhibit a substantial performance degradation, suggesting that pre-trained language models' contextual understanding generalizes more effectively than task-specific fine-tuning or fixed embeddings.

3. Datasets

Our study evaluates model performance across two datasets: (1) an in-distribution (ID) dataset with keyword-filtered CVEs for training and testing, and (2) an out-of-distribution (OOD) dataset with unfiltered recent CVEs to assess robustness under concept drift.

3.1. In-Distribution(ID) dataset: 5G-CVE-DB

We utilize a balanced subset of 510 CVEs from a publicly available dataset [23] containing 1,531 manually labeled network infrastructure vulnerabilities. The parent dataset was constructed by applying keyword-based filtering, as an example (*"5G", "gNB", "AMF", "NGAP"*), to CVE records published by NIST NVD between January 1, 2019 - July 1, 2025 followed by manual annotation based on semantic clustering of CVEs[24].

Our balanced subset, referred to as *5G-CVE-DB* throughout this paper, contains 255 5G-relevant and 255 non-relevant to 5G CVEs. Each record includes: CVE ID, Common Platform Enumeration (CPE), vulnerability descriptions, CVSS scores, and vectors, Common Weakness Enumeration (CWE) identifiers, and binary 5G-relevance labels. The relatively small corpus of 5G-related vulnerabilities reflects a systemic characteristic of the domain rather than a limitation of our data collection methodology. Currently, there are no publicly available CVE databases specifically specialized in 5G security vulnerabilities. Even proprietary commercial databases contain a comparable order of magnitude of 5G-relevant CVEs.

Selection Bias. *5G-CVE-DB* is obtained by keyword filtering on NIST NVD Dataset. This introduces a significant selection bias, leading to optimistic performance estimates that do not generalize to unfiltered deployment scenarios. This constitutes a *distribution shift* where $P_{\text{train}}(X) \neq P_{\text{deploy}}(X)$, meaning that the statistical properties of the data distribution change between training and deployment. Specifically, in our context, the deployment phase involves unfiltered CVEs from other technological sectors, potentially causing a mismatch between the 5G-focused data the model was trained on and the different types of vulnerabilities it may encounter. Combined with *concept drift*, this creates a dual challenge since the model faces both temporal and spatial shifts due to deployment on unfiltered data.

3.2. Out-Of-Distribution(OOD) dataset: Unfiltered Recent CVEs

To assess model robustness under both *concept drift* and *distribution shift*, we collected 9,186 CVE records published between July 1 and September 13, 2025, beyond the temporal scope of training data (ID). Critically, this dataset is completely unfiltered with no keyword pre-selection, so it represents a realistic challenge of processing newly published CVEs where 5G-relevant CVEs constitute a small minority.

Due to annotation cost, we employed a two-step approach for finding 5G related CVEs: (1) The embeddings-based SVM trained on *5G-CVE-DB* evaluated all 9,186 records, flagging 123 as potentially 5G-relevant; (2) Security experts manually validated these 123 candidates by analyzing CVE descriptions, CPE entries, and technical specifications. The final OOD dataset contains 31 5G CVEs and 92 no5g CVEs;

4. Methodology

Our methodology empirically compares four paradigms for 5G CVE classification: (1) embedding-based SVM, (2) large-scale 32B LLM model, (3) quantized 4B baselines LLM models, and (4) *LoRA*-adapted 4B LLM models with knowledge distillation. For the experiments, the Qwen3 model family was chosen due to its reasoning capabilities, computational efficiency, and quantization robustness[25]. For computational efficiency, we utilize *Qwen3-4B* with 4-bit quantization (*Q4_K_M format*), which reduces the model to approximately 2.5GB. Throughout this paper, we denote this variant as *Qwen3-4B (4bit)*.

Using the same prompt for the LLMs, we then evaluate each paradigm across ID and OOD datasets, testing whether the models can handle temporal *concept drift* and *distribution shifts* in vulnerability terminology to answer the following research questions:

- RQ1: Do larger language models (32B parameters) outperform smaller quantized models (4B parameters) on specialized binary classification tasks?
- RQ2: Does task-specific fine-tuning improve out-of-distribution classification performance compared to zero-shot inference ?
- RQ3: Do fine-tuned LLM classifiers achieve higher robustness than SVM embeddings when evaluated on out-of-distribution datasets?

Finally, we demonstrate a practical deployment through a lightweight prototype system that integrates the fine-tuned classifier with the NVD API for real-time vulnerability assessment.

4.1. Embedding-based SVM-RBF

We employ an ML pipeline where CVE descriptions and CPEs from *5G-CVE-DB* are concatenated and encoded using `all-MiniLM-L12-v2` [11]. These dense embeddings, unlike bag-of-words or TF-IDF representations, capture semantic similarity, so CVEs with similar technical content receive similar representations even without showing a lexical overlap. To determine optimal hyperparameters for the SVM classifier, we performed grid search with 5-fold cross-validation exclusively on the training set (408 samples). The search space included a regularization parameter $C \in \{0.1, 1, 2, 10, 100\}$ and RBF kernel coefficient $\gamma \in \{0.001, 0.01, 0.1, 1\}$.

For each hyperparameter configuration, the training set was divided into 5 folds, with the model trained on 4 folds (327 samples) and validated on the remaining fold (81 samples). This process was repeated 5 times with different held-out folds, and the mean validation accuracy across folds was computed. The final model was trained with the hyperparameters ($C = 1$, $\gamma = scale$, and a kernel RBF) with the highest mean cross-validation accuracy.

4.2. Prompt and LLM baselines

The prompt for LLMs, as shown in Listing 1, starts with role assignment, designating the model as a telecommunications security analyst, and defining the primary prompt task of binary classification. Then, the classification criteria for two classes are explained: positive indicators of 5G relevance, such as specific network elements, protocols, and deployment contexts, and negative indicators, including legacy technologies, consumer devices, and false-positive patterns.

The procedural workflow outlines five sequential steps, ranging from CVE description interpretation to CPE examination and with a final classification with a structured JSON output format containing the LLM reasoning, useful for debugging and explainability, and the LLM classification label as relevant or not to 5G (5G/no5G).

```
You are a telecommunications security analyst specializing in 5G infrastructure. Your task is
    to determine whether a CVE is relevant to 5G networks.
You are provided with a CVE summary and, if available, a list of CPEs (Common Platform
    Enumerations) identifying the affected products.
Your classification must be **strict and binary**, using only:
"5g" if the vulnerability directly or indirectly affects 5G networks, components, or protocols.
"no5g" if it is unrelated to 5G, or belongs to legacy tech like 4G, 3G, or unrelated fields
    (e.g., 5GHz Wi-Fi, IoT, industrial devices).

### Classification Criteria
A CVE is **5G-related** if:
It affects 5G network elements, such as:
  - Radio Access Network (RAN), gNodeB, 5G Core, MEC
  - 5G UE, chipsets, 5G modems or antennas
It comprises any of the following **5G-specific protocols**:
  - NGAP, PFCP, SCTP, NAS, GTP-U, HTTP/2, RRC
  - Or Linux kernel modules that handle the above protocols
The affected product is confirmed (via CPE or web results) to be **deployed in 5G mobile
    networks**
A CVE is **not 5G-related** if:
It affects legacy tech (LTE, 4G EPC, 3G, etc.)
It affects non-5G consumer devices (e.g., Wi-Fi routers, home cameras, enterprise software)
The mention of "5G" is unrelated to telecom networks (e.g., "5Ghz Wi-Fi" 5Ghz frequency band,
    product model names)
### Steps
1. Read and interpret the CVE summary carefully.
2. Examine the CPEs to identify the product type and version.
3. Check whether the affected components or protocols are used in 5G networks.
4. Filter out false positives (e.g., "5GHz Wi-Fi", 5Ghz frequency band, older mobile
    generations).
5. Use only "5g" or "no5g" as the final classification label.
### Answer using the following JSON format
{{
  "reasoning": "Explain clearly why this CVE is or isn't related to 5G.",
  "classification": "5g | no5g"
}}
```

Listing 1: Prompt to build syntetic conversational dataset.

4.3. Training Data Augmentation and Fine-Tuning Strategy

Since *Qwen3-32B* has high computational requirements, we transfer its classification capabilities to a smaller *Qwen3-4B(4bit)* model by knowledge distillation in combination with parameter-efficient fine-tuning via *LoRA*.

We use *Qwen3-32B* as a teacher model to generate reasoning chains for training samples in the *5G-CVE-DB*. For each CVE, we prompt the teacher with the vulnerability description, affected CPE list, and ground-truth label. The teacher produces step-by-step reasoning explaining why the vulnerability

is or is not relevant to 5G, as shown in Table 1, where, starting from the CVE Description, the CPE, and the correct label, the teacher motivates the choice by reasoning.

Table 1
Qwen3-32B Reasoning for in-distribution data (ID), data that comes from the same distribution on which the model was trained

CVE ID	CVE Description	CPE	5G/no5G	Qwen3-32B Reasoning
CVE-2024-24428	A reachable assertion in the oai_nas_5gmm_decode function of Open5GS <= 2.6.4 allows attackers to cause a Denial of Service (DoS) via a crafted NGAP packet.	cpe:2.3:a:open5gs:open5gs	5g	The vulnerability affects the oai_nas_5gmm_decode function in Open5GS, which is a 5G core network implementation. This function is part of the 5G Non-Access Stratum (NAS) protocol handling, specifically for the 5G Mobility Management (MM) messages. Since it involves a 5G-specific protocol and is part of the 5G core stack, this CVE is directly related to 5G.
CVE-2024-6206	A security vulnerability has been identified in HPE Athonet Mobile Core software. The core application contains a code injection vulnerability where a threat actor could execute arbitrary commands with the privileges of the underlying container, leading to complete takeover of the target system.	-	5g	The HPE Athonet Mobile Core software is part of the 5G Core network infrastructure. The vulnerability allows for code injection and system takeover, which could directly impact the integrity and operation of a 5G mobile core network. Since the affected product is deployed in 5G mobile networks and is integral to 5G functionality, this CVE is classified as 5G-related.

During distillation, the student model learns from both the teacher's final classifications and its synthetic reasoning process. Instead of fine-tuning all 4 billion parameters, we employ LoRA, which freezes the original pre-trained weights while injecting trainable low-rank matrices into the attention projection layers. This approach drastically reduces trainable parameters and memory requirements, making fine-tuning feasible on free-tier GPUs like Google Colab.

Following the same 80/20 split previously used for the embedding-based *SVM-RBF*, we trained the student model using Unsloth [26], an open source framework for efficient LLM fine-tuning.

To investigate the *LoRA* hyperparameters' impact on classification performances, we then tested three setups following standard Unsloth guidelines. We varied two parameters: *rank* (r), which sets how large the added matrices are and therefore how many parameters get trained, and *alpha* (α), which controls how much weight the *LoRA* updates get compared to the original model as follows:

- r8/α8: Baseline configuration with α = r, following Unsloth's recommended 1:1 ratio for stable fine-tuning.
- r8/α32: Higher scaling with α = 4r to test stronger *LoRA* influence while keeping rank low.
- r16/α32: Standard 2:1 ratio (α = 2r) with increased rank to add model capacity.

All other training hyperparameters were held constant across configurations (*learning rate*, *batch size*, *epochs*) to isolate the effects of *rank* and *alpha* scaling.

4.4. Implementation Details

The training and classification experiments were conducted with multiple hardware configurations depending on the computational requirements of the corresponding pipeline stages. We used a MacBook Pro M2 Pro for compact LLM models, an RTX 4090 workstation for 32B LLM, and Google Colab for fine-tuning smaller LLM with the Unsloth library. Hardware specifications are detailed in Table 2.

Configuration	CPU	RAM	GPU	VRAM	TFLOPs (FP32)
m2_pro	M2 Pro	16 GB	Integrated	16 GB*	5.68
rtx_4090	Ryzen 9	64 GB	RTX 4090	24 GB	82.58
colab	Virtualized	12.7 GB	Tesla T4	15 GB	8.1

Table 2
Hardware configuration used for experiments

5. Results

5.1. Evaluation on 5G-CVE-DB (in-distribution data)

Table 3 reports the results obtained by the *SVM-RBF*, the fine-tuned models, *Qwen3-4B(4bit)* and *Qwen3-32B* on 5G-CVE-DB, the in-distribution dataset. The *SVM-RBF* with embeddings achieves an F1 score of 0.92, closely matching the LLMs' more complex approaches. Given the limited test set size and absence of formal significance testing, these differences are insufficient to establish definitive model superiority. However, this finding prompts the question of whether LLMs are necessary for this task, while the *SVM-RBF* obtains comparable performances. Furthermore, the *Qwen3-32B* model does not outperform the others, achieving perfect precision (1.00) but lower recall (0.84), suggesting a model tendency toward excessive caution in classification. On the other hand, the *LoRA* experiments indicate that hyperparameter tuning has a greater impact than model size: increasing alpha from 8 to 32 yields the highest F1 score (0.93), whereas doubling the rank to 16 reduces performance (F1 0.88), likely due to overfitting.

Model	Accuracy	Precision	Recall	F1 Score
SVM-RBF	0.9118	0.92	0.902	0.92
Qwen3-4B(4bit)	0.92	0.89	0.96	0.92
Qwen3-4B(4bit) r8 la8	0.92	0.94	0.90	0.92
Qwen3-4B(4bit) r8 la32	0.93	0.96	0.90	0.93
Qwen3-4B(4bit) r16 la32	0.89	0.93	0.84	0.88
Qwen3-32B	0.92	1.00	0.84	0.91

Table 3
Performance of fine-tuned models on in-distribution dataset (5G-DB-CVE)

5.2. Evaluation under temporal shift and conceptual drift (out-of-distribution data)

Table 4 shows the results of classification on the OOD dataset, which contains 31 true 5G-relevant vulnerabilities and 92 no 5G vulnerabilities. The overall results of the classifiers are then presented in Table 4. It is noted that while *SVM-RBF* precision degrades on OOD data, *Qwen3-4B(4bit)* maintains similar or better performances with respect to distilled models as well as the *Qwen3-32B* version.

Model	Precision
SVM-RBF	0.244
Qwen3-4B(4bit)	0.91
Qwen3-4B(4bit) finetuned r8 la8	0.86
Qwen3-4B(4bit) finetuned r8 la32	0.85
Qwen3-4B(4bit) finetuned r16 la32	0.9
Qwen3-32B	0.91

Table 4
Performance comparison on 123 CVE records flagged as 5G-related by SVM on out-of-distribution dataset (9,186 published CVEs, July-September 2025).

This finding suggests a limitation of embedding-based SVM classifiers in this context, as their decision

boundaries depend primarily on static semantic representations derived from the training data. So, the SVM-RBF achieves competitive in-distribution performance (F1 0.92), but fails on out-of-distribution CVEs with a 0.24 precision versus an average 0.91 from LLMs approaches. When exposed to broader CVE records with evolved terminology or shifted contextual patterns, fixed embeddings seem to fail to capture evolved semantic relationships, leading to high false-positive rates.

A notable example is the record CVE-2025-53816, describing a vulnerability in *7-Zip*, a software clearly unrelated to 5G technology, with the following CVE description: *"7-Zip is a file archiver with a high compression ratio. Zeroes written outside heap buffer in RAR5 handler may lead to memory corruption and denial of service in versions of 7-Zip prior to 25.0.0."*. The SVM incorrectly classified this vulnerability as 5G-related. The semantic embeddings used (*all-MiniLM-L12-v2*) should distinguish between *"RAR5"* (a compression format) and telecommunication terms, but the SVM's linear decision boundary appears to have learned spurious correlations during training on the limited dataset of 510 CVEs. When the same description was slightly modified by masking the term *RAR5*, the model correctly reclassified the CVE as non-5G, suggesting that the SVM relies on specific token presence rather than semantic understanding.

Even the Qwen3-32B model exhibits some incorrect classifications on the OOD dataset. For CVE-2025-27072 it seems the model proceeds to an overgeneralization with superficial associations as "*Qualcomm equals mobile baseband then equals 5G*" so a vulnerability in EAVB packet, Ethernet Audio Video Bridging, an automotive protocol for synchronized media in vehicles, is associated to 5G with the following reasoning "*Qualcomm chipsets are likely baseband processors responsible for 5G connectivity*".

Moreover, the evaluation of the knowledge transfer points out that distilling may produce negative transfer with students who inherit poor decision boundaries, amplifying problematic behaviors, rather than learning generalizable patterns. We observed that this negative transfer manifests in two failure modes: hallucination and conservatism. CVE-2025-45619 affects the Aver PTC310UV2, a video conferencing camera that both quantized 4B baseline and 32B correctly identify as non-5G. The student LLM instead misclassifies it as 5G, claiming the device is "*a 5G baseband processor or modem*" which is untrue, suggesting spurious pattern learning from limited training data. Second, the negative transfer seems to amplify the 32B teacher's conservative approach. The CVE-2025-21427 describes RTP vulnerabilities in mobile User Equipment that both quantized 4B baseline and 32B were correctly classified as 5G-relevant. The distilled student rejects it, motivating it "*does not directly affect 5G-specific components.*". The student seems to have internalized the teacher's preference for explicit 5G mentions so rigidly that it now rejects cases even the 32B teacher accepts, becoming more restrictive than its source.

Overall results suggest that for well-defined binary tasks, as distinguishing 5G from non 5G CVEs, compact architectures as *Qwen3-4B(4bit)* may be advantaged with four-bit quantization, apparently acting as a form of regularization.

5.3. Limitation

The usage of OOD data is constrained by the SVM-based pre-selection process since the models were evaluated only on the 123 records initially flagged by *SVM-RBF*. Consequently, the classification on OOD measures precision but lacks comprehensive performance metrics such as Recall, Accuracy, and F1-Score, which are fully reported for the ID dataset. This limitation prevents a direct comparison of model degradation under concept drift and may obscure the presence of false negatives—a critical concern in vulnerability-detection contexts, where missed CVEs pose substantial security risks. The recall on the full corpus remains unknown, as the remaining 9,063 records were not manually annotated, meaning that 5G-relevant CVEs potentially missed by the SVM-RBF pre-selection could not be detected. While precision on the pre-selected subset provides initial insights into model behavior on OOD data, a more rigorous evaluation would require a comprehensive assessment with representative random sampling from the entire corpus. Future work should address these limitations by: (1) implementing a random sampling of the full OOD corpus to enable calculation of all performance metrics (2) providing a unified results table comparing ID and OOD performance across all metrics to quantify the exact degradation under distribution shift and (3) conducting end-to-end evaluation of the complete classi-

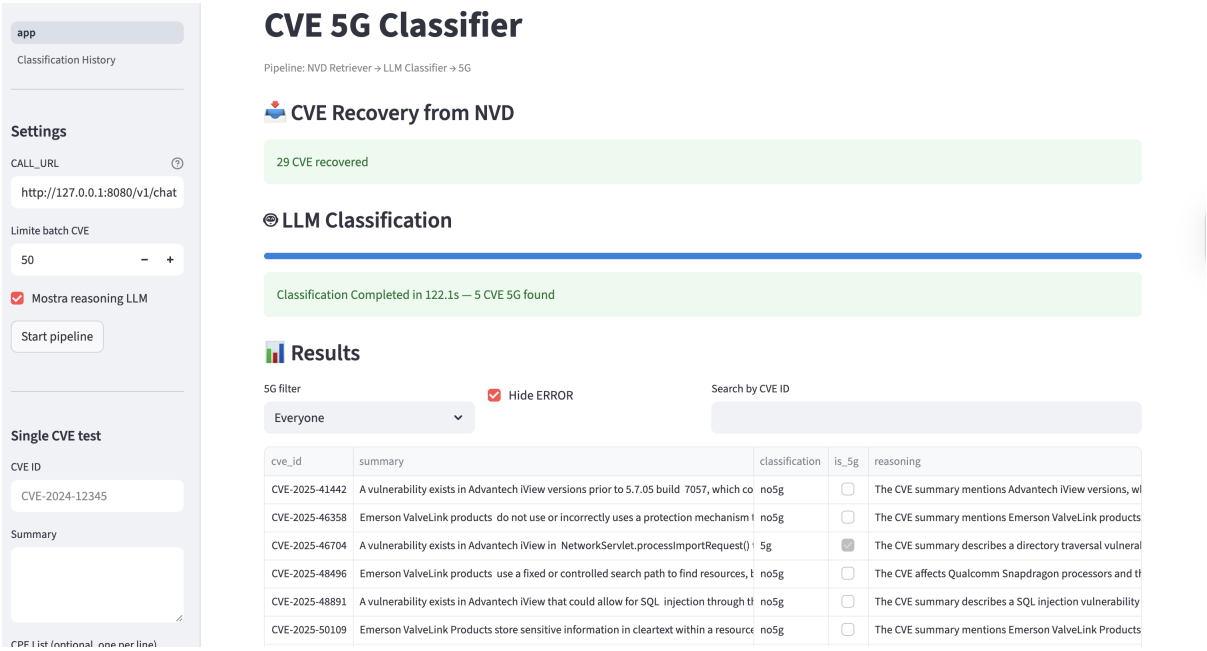

Figure 1: 5G CVE prototype classifier

fication pipeline to assess both the SVM pre-selection and the fine-tuned models' ability to identify 5G-relevant vulnerabilities across the entire dataset.

5.4. Prototype Development

To demonstrate the applicability of the proposed classification approach, a lightweight prototype system was developed as shown in Fig. 1. The system retrieves CVE records from the NVD API, extracts the fields used by the classifier, such as CVE description and CPEs, and provides a prediction if the CVE is related to 5G or not, together with a concise reasoning generated by the *Qwen3-4B(4bit)* in Table 5.

6. Conclusions

This work studies classifiers for facing the phenomenon of concept drift for CVEs related to 5G, where the technology and the language describing vulnerabilities evolve, as newer CVEs related to different components and attack vectors emerge.

Our comparative evaluation examines four classification strategies on a manually annotated training dataset of 510 5G CVEs (2019-2025) and on a temporally separated and more recent out-of-distribution dataset. Results show that semantic embeddings-based *SVM-RBF* cannot capture evolving 5G terminology (*concept drift*) or generalize from explicit keyword-rich training examples to implicit relevance cases on the out-of-distribution dataset (*distribution shift* due to selection bias of the training dataset).

In contrast, large language models maintain robust performance on out-of-distribution data. Critically, *Qwen3-4B(4bit)* matches *Qwen3-32B* despite being 8 times smaller, demonstrating that appropriate capacity is sufficient for this specialized binary task.

Moreover, distilling reasoning chains from the *Qwen3-32B* teacher to the *Qwen3-4B(4bit)* student using *LoRA* on keyword-filtered training data degrades out-of-distribution performance, introducing two failure modes: hallucination (creating spurious 5G associations) and over-conservatism (rejecting valid 5G CVEs due to over-reliance on explicit terminology). This suggests that fine-tuning may lead to negative transfer, as in the case of a selection bias on the training dataset, whereas base models' broader pre-training provides, even for a small 4B model, more robust semantic understanding.

This work demonstrates that compact, explainable LLMs can provide a practical solution for automated

Table 5
Reasoning for out-of-distribution data (OOD), data that has different characteristics compared to the training set

CVE ID	CVE Description	CPE	5G/no5G	Qwen3-4B(4bit) reasoning
CVE-2025-29525	DASAN GPON ONU H660WM OS version H660WMR210825 Hardware version DS-E5-583-A1 was discovered to contain insecure default credentials in the modem's control panel.	-	no5g	The vulnerability affects the GPON ONU H660WM, which is a component of the Optical Transport Network (OTN) and not directly related to 5G infrastructure. The issue lies in insecure default credentials for the modem's control panel, which is unrelated to 5G-specific components or protocols.
CVE-2025-32100	An issue was discovered in Samsung Mobile Processor, Wearable Processor, and Modem Exynos 980, 990, 850, 1080, 2100, 1280, 2200, 1330, 1380, 1480, 2400, 1580, 9110, W920, W930, W1000, Modem 5123, Modem 5300, and Modem 5400. A programming mistake for buffer copy leads to out-of-bounds writes via malformed ROHC packets.	cpe:2.3:h:samsung:exynos 1080/1280/1330/1380/1480/1580/2100/2200/2400/850	5g	The CVE affects Samsung's Exynos processors and modems, which are used in devices that support 5G communication. The vulnerability is related to the handling of ROHC (Robust Header Compression) packets, which are part of the 5G New Radio (NR) access network protocols. Since the issue pertains to the baseband processor and modem, which are critical 5G components in user equipment, this CVE is relevant to 5G networks.

domain-specific vulnerability classification. By leveraging detailed prompt engineering, these models generate structured reasoning chains that make classification decisions transparent and auditable, a critical capability absent in traditional approaches. The comparison between LLMs and traditional classifiers should not be framed solely on performance metrics or computational costs, but rather on their operational value in real-world deployment scenarios. The dynamic nature of 5G vulnerabilities, with continuously evolving standards, emerging attack vectors, and shifting vendor terminology, necessitates a human-in-the-loop evaluation process where security analysts make final classification decisions. In this context, the primary value proposition of LLM-based classifiers lies not in fully autonomous classification but in providing explainable decision support that enhances analyst productivity and decision quality.

Acknowledgments

This work was partially funded by the project SERICS (PE00000014) under the NRRP MUR program funded by the EU - NGEU.

Declaration on Generative AI

During the preparation of this work, the authors employed the tool Grammarly to ensure grammatical accuracy, correct spelling errors, and refine phrasing for enhanced readability. After using this tool, the authors reviewed and edited the content as needed and took full responsibility for the publication's content.

References

[1] Éireann Leverett. *Vulnerability Forecast for 2025*. https://www.first.org/blog/20250607-Vulnerability-Forecast-for-2025. Accessed: 10 December 2025. 2025.

[2] VexGen Project. *NVD CVE Analysis Rate Report*. https://vexgen.github.io/. Accessed: 10 December 2025. 2025.

[3] Kobra Khanmohammadi and Raphaël Khoury. "Half-day vulnerabilities: a study of the first days of CVE entries". In: *arXiv preprint arXiv:2303.07990* (2023).

[4] Mohammed Khalil. *Vulnerabilities Statistics 2025: CVE Surge & Exploit Speed*. Blog post, DeepStrike. Accessed: 10 December 2025. 2025. URL: https://deepstrike.io/blog/vulnerability-statistics-2025.

[5] João Gama et al. "A survey on concept drift adaptation". In: *ACM computing surveys (CSUR)* 46.4 (2014), pp. 1–37.

[6] Jie Lu et al. "Learning under concept drift: A review". In: *IEEE transactions on knowledge and data engineering* 31.12 (2018), pp. 2346–2363.

[7] Abdul Razak MS et al. "A survey on detecting healthcare concept drift in AI/ML models from a finance perspective". In: *Frontiers in Artificial Intelligence* 5 (2023), p. 955314.

[8] Andrea Dal Pozzolo et al. "Credit card fraud detection and concept-drift adaptation with delayed supervised information". In: *2015 International Joint Conference on Neural Networks (IJCNN)*. IEEE. 2015, pp. 1–8.

[9] Nicolas Jourdan et al. "Handling concept drift in deep learning applications for process monitoring". In: *Procedia CIRP* 120 (2023), pp. 33–38.

[10] Triet Huynh Minh Le, Bushra Sabir, and Muhammad Ali Babar. "Automated software vulnerability assessment with concept drift". In: *2019 IEEE/ACM 16th International Conference on Mining Software Repositories (MSR)*. IEEE. 2019, pp. 371–382.

[11] Nils Reimers and Iryna Gurevych. "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 3982–3992.

[12] Marios Kokkodis, Richard Demsyn-Jones, and Vijay Raghavan. "Beyond the Hype: Embeddings vs. Prompting for Multiclass Classification Tasks". In: *arXiv preprint arXiv:2504.04277* (2025).

[13] Niklas Muennighoff et al. "Mteb: Massive text embedding benchmark". In: *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. 2023, pp. 2014–2037.

[14] Martin Juan José Bucher and Marco Martini. "Fine-Tuned 'Small' LLMs (Still) Significantly Outperform Zero-Shot Generative AI Models in Text Classification". In: *arXiv preprint arXiv:2406.08660* (2024).

[15] Yu Fei et al. "Beyond Prompting: Making Pre-Trained Language Models Better Zero-Shot Learners by Clustering Representations". In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. 2022, pp. 8560–8579.

[16] Pierpaolo Bene et al. "Optimizing Local LLM Deployment for 5G CVE Classification Avoiding External Data Exposure". In: *2025 IEEE Conference on Communications and Network Security (CNS)*. IEEE. 2025, pp. 1–3.

[17] Rikhiya Ghosh et al. "Cve-llm: Ontology-assisted automatic vulnerability evaluation using large language models". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 39. 28. 2025, pp. 28757–28765.

[18] Francesco Marchiori, Denis Donadel, and Mauro Conti. “Can LLMs Classify CVEs? Investigating LLMs Capabilities in Computing CVSS Vectors”. In: *arXiv preprint arXiv:2504.10713* (2025).

[19] Jie Zhang et al. “When llms meet cybersecurity: A systematic literature review”. In: *Cybersecurity* 8.1 (2025), p. 55.

[20] Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, et al. “Distilling step-by-step! Outperforming larger language models with less training data and smaller model sizes”. In: *arXiv preprint arXiv:2305.02301* (2023).

[21] Wei Liu, Weihao Tang, Keqing Lu, et al. “What makes good data for alignment? A comprehensive study of automatic data selection in instruction tuning”. In: *arXiv preprint arXiv:2312.15685* (2024).

[22] Edward J Hu et al. “LoRA: Low-Rank Adaptation of Large Language Models”. In: *International Conference on Learning Representations (ICLR)*. 2022.

[23] Francesco D’Alterio, Andrea Bernardini, and Leonardo Sagratella. *5G and Network Infrastructure CVE-Annotated Dataset: Distinguishing 5G Native, LTE, Auxiliary to 5G, and Non-5G Vulnerabilities*. Data set. 2025. DOI: 10.5281/zenodo.17450053. URL: https://doi.org/10.5281/zenodo.17450053.

[24] Francesco D’Alterio et al. *Descriptor: A CVE Dataset for 5G and Related Network Infrastructure*. Dec. 2025. DOI: 10.36227/techrxiv.176538348.85407432/v1. URL: https://doi.org/10.36227/techrxiv.176538348.85407432/v1.

[25] Qwen Team. “Qwen3 Technical Report”. In: *arXiv preprint arXiv:2505.09388* (May 2025). URL: https://arxiv.org/abs/2505.09388.

[26] Michael Han Daniel Han. *Unsloth Documentation*. https://docs.unsloth.ai/. Accessed: 10 December 2025. 2025.