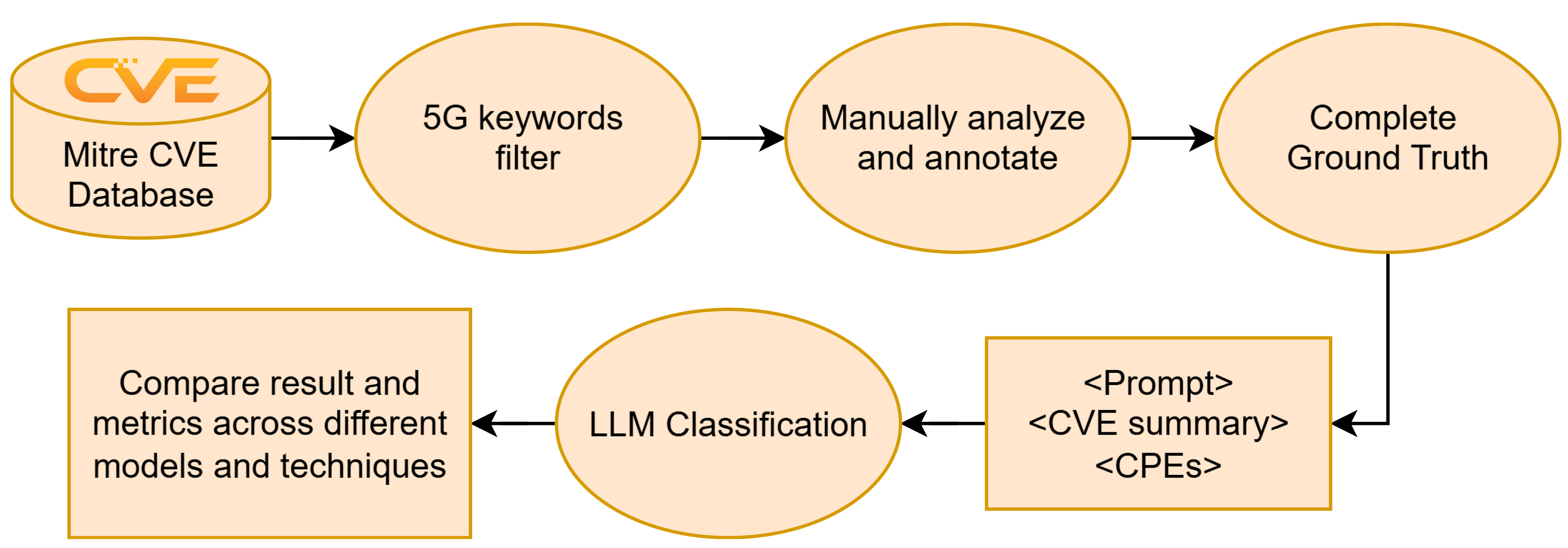
Optimizing Local LLM Deployment for 5G CVE Classification Avoiding External Data Exposure

Pierpaolo Bene[1], Andrea Bernardini[2], Leonardo Sagratella[2], Nicolo Maunero[3], Marina Settembre[2]
[1] Department of Control and Computer Engineering, Politecnico di Torino, Turin, Italy
[2] Fondazione Ugo Bordoni, Rome, Italy
[3] IMT School for Advanced Studies, Lucca, Italy

Motivation: The rapid growth of CVEs—projected to exceed 50,000 new entries in 2025—creates a major challenge for timely vulnerability management. While 5G-specific CVEs are still emerging, their complexity demands specialized expertise and rapid identification. Traditional methods like keyword filtering and manual review are too slow and error-prone to keep up. An automated, domain-aware solution is needed to classify 5G-related vulnerabilities as soon as they are published, without exposing sensitive data outside the organization.

Approach: We dataset and systematically tested locbuilt a manually annotated 5G-specific CVE al large language models (LLMs) for automated classification. Our evaluation progressed from simple baselines to advanced prompt-engineering strategies, including few-shot learning, context enrichment (via embeddings), and reasoning-based approaches. This enables efficient, privacy-preserving classification that leverages LLMs' natural language understanding and cross-domain knowledge.

1. Pipeline



2. Ground Truth

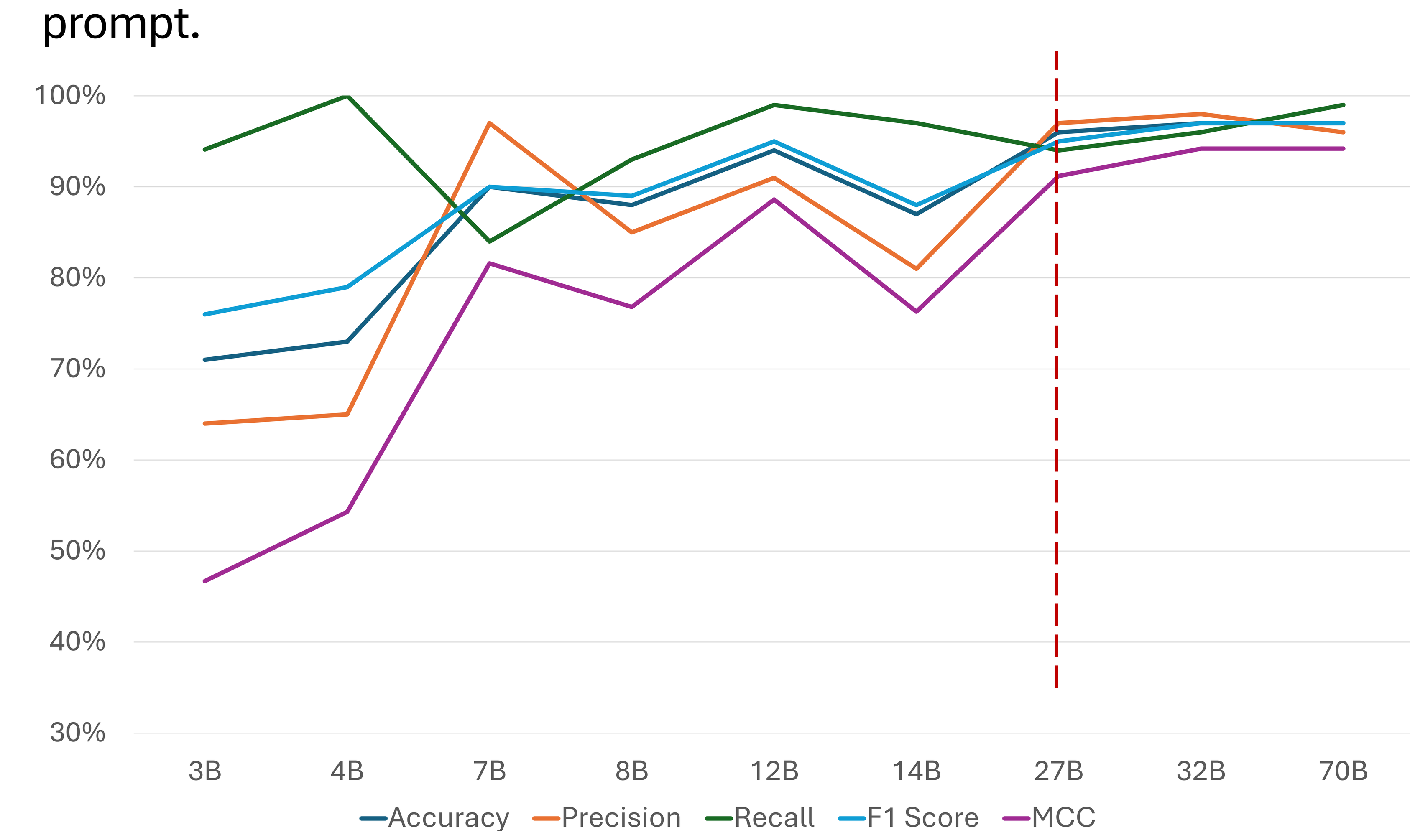
- Filter CVEs based on 5g keywords
- **136** CVE, manually annotated by three domain experts, covering 2014–2024.
- Binary classification:
  - **5g:** Vulnerabilities affecting 5G core network functions, RAN components, or 5G-specific protocols.
  - **no5g:** Vulnerabilities related to general networks, applications, or infrastructure not specific to 5G.
- Ground Truth is available at Zenodo, scan the QR code. ->

3. LLM evaluation

- Run multiple local LLM with different size (3B-70B).
- Different prompt engineering techniques:
  - Baseline (**B**): uses only CVE description and CPEs if present.
  - Few-shots (**FS**): provides two output example to the baseline prompt.
  - Web context enrichment using LLM (**CL**) or embeddings (**CE**): Enriches the prompt context with information gathered from CVE associated references. Useful information are summarized using either an embedding model or the LLM itself.
  - Reasoning or CoT (**R**): Asks the model to make some reasoning before giving the answer or enables reasoning mode when available.
- Evaluated *Accuracy*, *Precision*, *Recall*, *F1-score* and *Mattews Correlation Coefficient* (MCC) across different models and techniques. MCC is particularly suited for binary classification since it balances true/false positives and negatives.

4. Results

- Baseline metrics improvement with increasing parameter size.
- Performance plateau observed beyond ~14B parameters, with limited gains from scaling.
- Steady recall across the models.
- No significant improvement using different prompt engineering techniques with higher parameter size.
- Embedding-based enrichment is especially effective for small and mid-size models.
- Prompt sensitivity, results may vary significantly by changing the prompt.



Baseline strategy, metrics by model size.

Model	B (%)	FS	CL	CE	R
Llama-3B-Q4	46.7	+0.7	+3.0	**+12.9**	+6.9
Gemma-E4B-Q4	54.3	**+9.1**	+2.3	+5.6	+7.8
Mistral-7B-Q4	81.6	+6.1	+2.7	**+6.7**	-3.55
Llama-8B-Q4	76.8	+2.1	+5.7	**+8.8**	+1.6
Gemma-3-12B	88.6	-2.4	-6.2	-3.9	**-1.8**
SecGPT-14B	76.0	**+12.6**	+0.6	+7.2	+9.1
Gemma-27B-Q4	91.2	-7.9	+0.3	**+4.4**	+3.0
Qwen3-32B-Q4	94.2	+0.0	**+1.4**	-3.1	+0.1
Llama3-70B	94.2	-2.9	-2.9	-4.3	**+1.4**

MCC metric by model and strategy (in bold the highest increase per row).

5. Future Works

- Expand dataset and refine annotations (more categories).
- Explore fine-tuning of LLMs on 5G CVEs.
- Mitigate prompt sensitivity and explore robustness of quantization/temperature.
- More comprehensive model performance analysis.

References

[1] J. Leverett, "Vulnerability Forecast for 2025," FIRST.org Blog, Jun. 2025.
[2] A. K. Alnaim, "Securing 5G virtual networks: a critical analysis of SDN, NFV, and network slicing security," International Journal of Information Security, 2024.
[3] F. D'Alterio, M. Rotunno, M. Settembre, A. Bernardini, L. Sagratella, G. Bianchi et al., "Navigating 5G security: Challenges and progresses on 5G security assurance and risk assessment," in 2024 AEIT International Annual Conference (AEIT), IEEE, Sep. 2024, pp. 1–6.
[4] VulnCheck, 2025 Q1 Trends in Vulnerability Exploitation, 2025.
[5] Y. Wu, M. Wen, Z. Yu, X. Guo, and H. Jin, "Effective vulnerable function identification based on CVE description empowered by large language models," in Proc. 39th IEEE/ACM Int. Conf. Automated Software Engineering (ASE), 2024, pp. 393–405.
[6] NVIDIA Corporation, "Applying generative AI for CVE analysis at an enterprise scale," NVIDIA Technical Blog, May 2024.
[7] A. Bernardini, L. Sagratella, and F. D'Alterio, 5G CVE Dataset, Zenodo, 2025. [Online]. Available: https://doi.org/10.5281/zenodo.16736495
[8] Hugging Face, "all-mpnet-base-v2 model," 2021.
[9] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in Text Summarization Branches Out, pp. 74–81, 2004.
[10] D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," BMC Genomics, vol. 21, no. 1, p. 6, 2020.
[11] Research Team, "RANsacked: Over 100 security flaws found in LTE and 5G network implementations," The Hacker News, Jan. 2025.

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