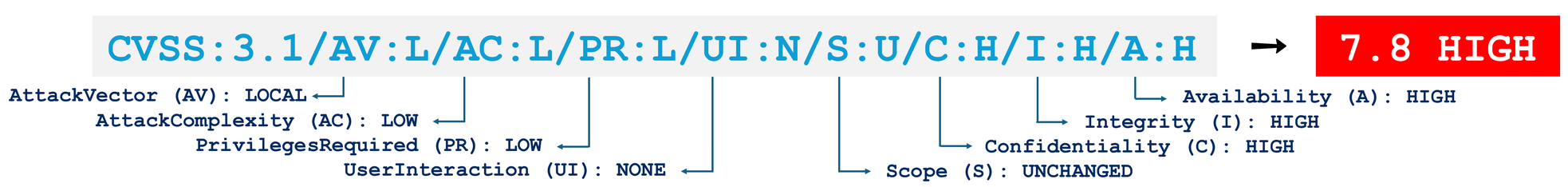
AutoCVSS: Assessing the Performance of LLMs for Automated Software Vulnerability Scoring

Davide Sanvito, Giovanni Arriciati, Giuseppe Siracusano, Roberto Bifulco, Michele Carminati

CVE and CVSS

- Common Vulnerabilities and Exposures (**CVE**) program
  - De-facto standard to identify and catalog publicly-disclosed **software vulnerabilities**
- Common Vulnerability Scoring System (**CVSS**) industrial standard
  - Methodology to compute **severity scores** for new vulnerabilities
  - Critical for vulnerability management, e.g., risk assessment and prioritization, incident response, etc.



The current cybersecurity landscape

- CVSS scores are *manually* assessed by National Vulnerability Database (**NVD**) analysts

Increased burden on cybersecurity analysts

- Growing number of disclosed vulnerabilities: +38% from 2023 to 2024 (40k)
- Median manual analysis delay increased from 2.7 days in 2019 to 8.2 in 2023

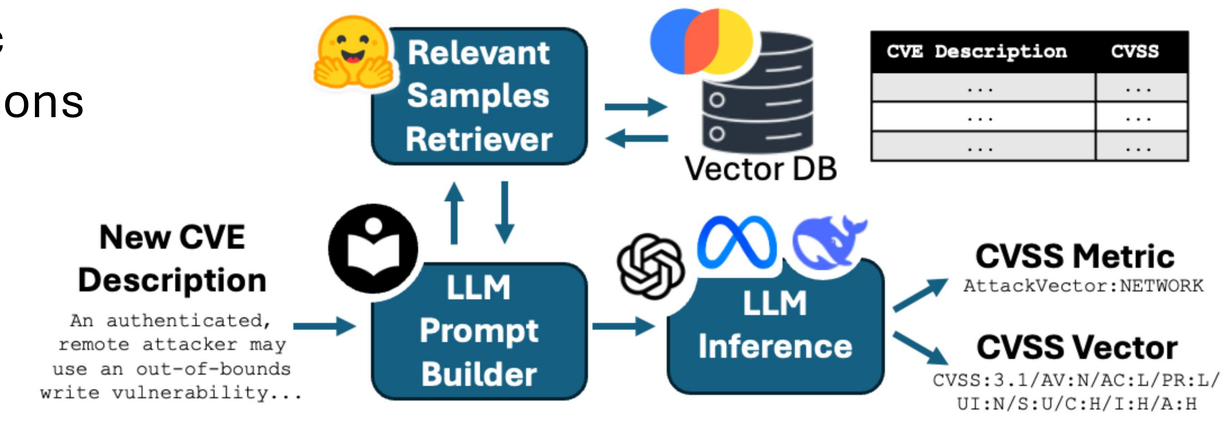
Attackers are moving faster

- Average time-to-exploit decreased from 44 days in 2019 to 5 days in 2023
- 25% of high-risk vulnerabilities is exploited on the day they are published

Larger window of opportunity for attackers!

AutoCVSS

- Assessment of LLM performance for the **automatic prediction of CVSS scores** based on CVE descriptions
- Multi-dimensional experimental analysis
  - Prompting: zero-shot vs few-shots with RAG
  - Models: open-source vs closed-source models
  - LLMs vs Large Reasoning Models (LRMs) variants
  - CVSS standard versions: v3.1 vs v4.0
  - Comparison against fine-tuned BERT supervised baselines (*DistilBERT-E*, *CVEDrill* and *ModernBERT*)



- AutoCVSS is open-source: https://github.com/nec-research/AutoCVSS

Results

- **Key takeaways**
  - **High data availability** (e.g. CVSS v3.1): LLMs offer competitive performance, but lag behind supervised approaches (Accuracy averaged on the CVSS metrics: 91.5% vs 92.9%)
  - **Low data availability** (e.g. transition from CVSS v3.1 to v4.0): LLMs surpass the supervised baselines for half of the CVSS metrics. A **hybrid approach** yields the best performance (92.2%)
  - **No data available** (e.g. new CVSS version released): LLMs represent the only viable approach, outperforming the worst-case conservative approaches (78%)
  - LLMs bridge the gap during early CVSS adoption and pave the way for more accurate hybrid and fully-supervised approaches as the availability of labelled data increases over time
- **Additional results** in the paper: full set of performance metrics, results break-down by CVSS metrics, CVSS Qualitative Severity Rating Scale (QSRS), LLM Knowledge Cutoff-aware evaluation, LLM Fine-tuning preliminary analysis

CODE

NEC POLITECNICO MILANO 1863 EMNLP 2025 Suzhou, China 中国苏州