Ofenzivna sigurnost

Primjena AI za penetracijska testiranja

Matea Teglović, 22.12.2025.

Pregled predavanja

- Motivacija
- Pitanja za ispite
- Penetracijsko testiranje i izazovi automatizacije
- Ograničenja klasičnih alata za ofenzivnu sigurnost
- Veliki jezični modeli (LLM)
- Primjena LLM-ova u penetracijskom testiranju (PentestGPT)
- End-to-end automatizirano penetracijsko testiranje
- Evaluacija, prednosti, ograničenja i rizici

Motivacija

- Porast složenosti sustava i napadačkih površina
- Penetracijsko testiranje zahtijeva visoku razinu stručnosti i značajno vrijeme
- Ograničenja klasičnih automatiziranih alata za penetracijsko testiranje
- Razvoj umjetne inteligencije i velikih jezičnih modela
- Potencijal umjetne inteligencije za automatizaciju i podršku ofenzivnim sigurnosnim testiranjima

Pitanja za ispite

- Koje su glavne prednosti korištenja velikih jezičnih modela u penetracijskom testiranju u odnosu na klasične automatizirane alate?
- Koju ulogu ima alat PentestGPT u automatizaciji penetracijskog testiranja?
- Što se podrazumijeva pod potpuno automatiziranim (end-to-end) penetracijskim testiranjem?
- U koju se svrhu koristi benchmark u evaluaciji alata za penetracijsko testiranje temeljenih na umjetnoj inteligenciji?
- Koja su ključna ograničenja i rizici primjene umjetne inteligencije u ofenzivnoj sigurnosti?

Penetracijsko testiranje (1)

- Kontrolirana simulacija stvarnih napada na sustav
- Cilj je otkriti i procijeniti sigurnosne ranjivosti
- Proces je iterativan i ovisi o rezultatima prethodnih koraka
- Naglasak je na donošenju odluka, ne samo izvršavanju testova

Penetracijsko testiranje (2)

- Problemi:
  - Odluke se donose na temelju nepotpunih informacija
  - Isti nalaz može imati više mogućih interpretacija
  - Napadi su često višekoračni i međusobno ovisni
  - Pogrešna odluka može uzrokovati ozbiljne posljedice

Kako se penetracijsko testiranje pokušava automatizirati?

- Ručno testiranje je vremenski i resursno zahtjevno
- U praksi se automatiziraju pojedine faze testiranja
- Automatizacija služi kao podrška, ne zamjena za pentestera
- Cilj je ubrzati otkrivanje potencijalnih problema

Ograničenja klasičnih automatiziranih alata

- Temelje se na statičnim pravilima i potpisima ranjivosti
- Ne razumiju širi kontekst ciljnog sustava
- Ne povezuju rezultate različitih alata
- Ne donose strateške odluke o daljnjim koracima

Što su veliki jezični modeli (LLM)?

- Modeli umjetne inteligencije trenirani na velikim količinama teksta
- Uče odnose, značenja i strukturu jezika
- Mogu razumjeti i generirati tehnički tekst
- Obraduju informacije, ali sami ne izvršavaju akcije

Veliki jezični model u ofenzivnoj sigurnosti

- Sudjeluje u procesu odlučivanja, ne samo analize
- Na temelju rezultata predlaže sljedeće korake
- Donosi odluke unutar definiranog cilja testiranja
- Ponaša se kao „inteligentni posrednik“ između alata

PentestGPT

- Sustav temeljen na velikom jezičnom modelu
- Djeluje kao koordinacijski sloj iznad alata
- Ne izvršava napade izravno
- Cilj je oponašati način razmišljanja pentestera

PentestGPT

- Tok rada:
  - Analizira izlaz postojećih sigurnosnih alata i skenera
  - Na temelju dobivenih informacija identificira moguće ranjivosti
  - Donosi odluku o najprikladnijem sljedećem koraku testiranja
  - Procjenjuje rezultat izvršene akcije i ažurira strategiju
  - Proces se ponavlja dok se ne iscrpe relevantni napadački vektori

Zašto PentestGPT nije klasična automatizacija?

- Ne koristi fiksna pravila ili potpise
- Odluke ovise o kontekstu i prethodnim rezultatima
- Proces nije linearan ni unaprijed definiran
- Sustav se prilagođava tijekom testiranja

Potpuno automatizirano (end-to-end) testiranje

- Automatizacija cijelog životnog ciklusa testiranja
- Od izviđanja do izvještavanja
- Minimalna ili nikakva ljudska intervencija
- Dugoročni istraživački cilj

Potpuno automatizirano (end-to-end) testiranje

- Problemi:
  - Rizik pogrešnih i nekontroliranih odluka
  - Nedostatak razumijevanja posljedica napada
  - Neočekivana ponašanja stvarnih sustava
  - Potreba za ljudskom odgovornošću

Evaluacija AI alata u penetracijskom testiranju

- Uspješan napad nije jedina metrika
- Važno je procijeniti proces odlučivanja
- Procjenjuje se konzistentnost i pouzdanost
- Potrebni su standardizirani kriteriji

Uloga benchmarka u evaluaciji AI alata

- Objektivna usporedba različitih pristupa
- Standardizirani sigurnosni scenariji
- Evaluacija procesa odlučivanja
- Procjena pouzdanosti i konzistentnosti
- Smjernice za daljnji razvoj sustava

Prednosti primjene AI-ja

- Skalabilnost sigurnosnih testova
- Brža analiza velikih količina podataka
- Podrška stručnjacima u odlučivanju
- Mogućnost kontinuiranog testiranja

Ograničenja i rizici

- Halucinacije i pogrešni zaključci
- Eskalacija grešaka kroz više koraka
- Nedostatak objašnjivosti odluka
- Potreba za ljudskim nadzorom

Sigurnosni i etički rizici

- Mogućnost zloupotrebe AI alata u stvarnim napadima
- Automatizacija potencijalno štetnih aktivnosti
- Odgovornost za posljedice korištenja alata

Zaključak

- Umjetna inteligencija unapređuje penetracijsko testiranje
- Najveća vrijednost je u podršci odlučivanju
- Ne može u potpunosti zamijeniti stručnjaka
- Najbolji pristup: AI + struktura + ljudski nadzor

Literatura

- **Isozaki, I., Shrestha, M., Console, R., Kim, E.**
  “Towards automated penetration testing: Introducing LLM benchmark, analysis, and improvements.” *UMAP / ACM*, 2025.
- **Happe, A., Cito, J.**
  “On the surprising efficacy of large language models for penetration testing.” *arXiv*, 2025.
- **Deng, G., et al.**
  “PentestGPT: An LLM-empowered automatic penetration testing tool.” *arXiv*, 2023.
- **Deng, G., et al.**
  “PentestGPT: Evaluating and harnessing large language models for automated penetration testing.” *USENIX Security Symposium*, 2024.
- **Wu, B., et al.**
  “AutoPT: How far are we from end-to-end automated web penetration testing?” *arXiv*, 2024.

Dodatna literatura

- **Nakatani, S.**
  “RapidPen: Fully automated IP-to-shell penetration testing with large language model–based agents.” *arXiv*, 2025.
- **Ginige, Y., et al.**
  “AutoPentester: An LLM agent-based framework for automated penetration testing.” *arXiv*, 2025.

Hvala!