



Imolai Dávid – Muzsai Lajos

Eötvös Loránd Tudományegyetem, Eötvös József Collegium

HackSynth - Nagy nyelvi modell alapú ügynök és kiértékelő keretrendszer autonóm sérülékenységteszteléshez



A kiberfenyegetések gyors növekedése igényt teremt robusztus és skálázható kiberbiztonsági megoldások iránt. A penetrációs tesztelés hagyományosan nagymértékben támaszkodik emberi szakértőkre, azonban a rendszerek bonyolultságának és a potenciális sérülékenységek számának növekedésével a manuális megközelítés egyre kevésbé praktikus.

Ezen problémák megoldására bemutatjuk a *HackSynth*-et, egy új, nagy nyelvi modell (*Large Language Model*, LLM) alapú ügynököt, amely képes autonóm sérülékenységtesztelésre. A *HackSynth* architektúrája két modul tartalmaz, a *Tervezőt* és az *Összegzőt*, melyek feladata parancsok generálása, illetve a kapott válaszok feldolgozása iteratív módon. A *HackSynth* sérülékenységtesztelési képességeinek értékeléséhez két új, Capture The Flag (CTF) feladatokból álló *benchmark*ot készítettünk, amelyek a *PicoCTF* és *OverTheWire* platformokon alapulnak. A két gyűjtemény együtt kétszáz CTF-feladatot tartalmaz, hat kategóriából és változatos nehézségi szintekből, szabványos keretrendszert biztosítva a teszteléshez.

Az új *benchmark*okat felhasználva részletes méréseken keresztül elemezzük a *HackSynth* olyan alapvető paramétereinek teljesítményre, hibázásra való hatását, mint a meghajtó LLM-kreativitásának szintje (*temperature* és *top-p*) vagy a válaszokból felhasznált tokenek mennyisége. Méréseinkkel több nyílt és zárt LLM-mel meghajtott ügynök képességeit hasonlítottuk össze. Kísérleteink során azt tapasztaltuk, hogy a *HackSynth* hatékonyan képes megoldani a CTF-feladatok jelentős részét. Biztonsági értékelést is végeztünk, amely kiemelte, hogy az autonóm ügynökök felhasználása során robusztus védelmeket kell alkalmazni a nem szándékos viselkedések megelőzésének érdekében.

Eredményeink azt mutatják, hogy az LLM-alapú autonóm penetrációs tesztelés egy ígéretes irány, emellett hangsúlyozzuk a kapcsolódó biztonsági intézkedések szükségességét is. A *HackSynth* ügynök és a két *benchmark* nyilvánosan elérhetőek, hogy segítsék az autonóm kiberbiztonsági megoldások kutatását.