CTI Jelentés

A ChatGPT kiberbiztonsági kockázatai

NEMZETBIZTONSÁGI SZAKSZOLGÁLAT

NEMZETI KIBERVÉDELMI INTÉZET

Tartalomjegyzék

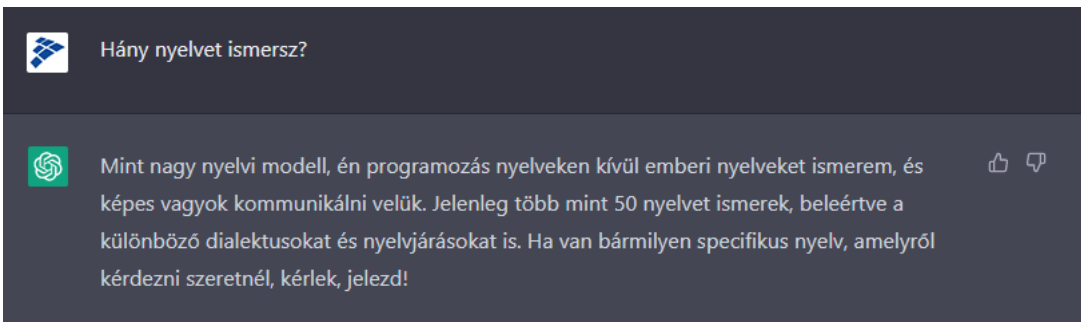
Mi az a ChatGPT?

A ChatGPT (Chat Generative Pre-trained Transformer) egy chatbot, amelyet az OpenAI kutatólaboratórium fejlesztett ki.

Általában chatbotként írják le, azonban ennél sokkal többre képes: szövegeket generál, összetett témákat magyaráz el, képes fordítani, vicceket kitalálni és kódot írni. Azonban a mesterséges intelligencia destruktív módon is felhasználható, amely új lehetőséget nyújt a kiberbűnözők számára.

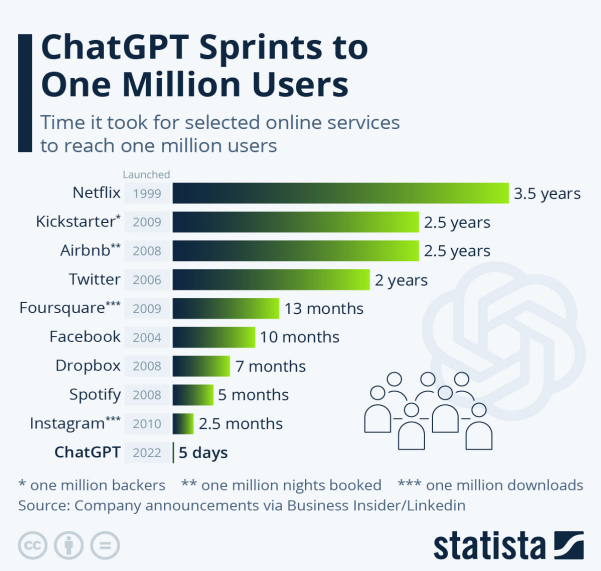
A GPT 3.5-ös verziója 2022. március 15-én jelent meg, a ChatGPT pedig ennek a verziónak a finomhangolásával jött létre és indult útjára 2022. november 30-án. A hírek szerint a GPT 4-es verziója már sokkal fejlettebb lesz, mint a jelenlegi ChatGPT, és 2023-ban már elérhetővé válik.

A szolgáltatás alapnyelve az angol, de néhány más nyelven, így például magyarul is működik. Állítása szerint pedig több, mint 50 nyelvet ismer, ezáltal meglehetősen sok nyelven is elmagyarázható neki, hogy mire van szükségünk.



A szolgáltatás a chat.openai.com oldalon érhető el a regisztrációt követően.

2022. december 4-én az OpenAI elnökének becslése szerint a ChatGPT-nek már több, mint egymillió felhasználója volt. Összehasonlításképpen a Twitternek két évébe telt, mire elérte az egymillió felhasználót. A Facebooknak tíz hónap, a Dropboxnak hét hónap, a Spotify-nak öt hónap, az Instagramnak pedig hat hét kellett.



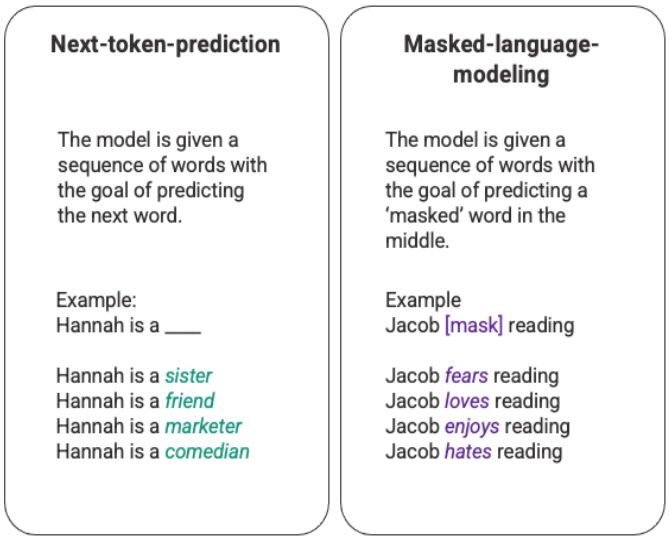
Forrás: www.statista.com

A szolgáltatást ingyenesen indították el a nagyközönség számára, azonban alig több, mint két hónappal később bejelentették a szolgáltatás fizetős változatának bevezetését.

A ChatGPT Plus kísérleti előfizetési csomag 2023. február 1-jétől érhető el havi 20 dollárért. Az OpenAI a díjért cserébe többek között azt az ígéretet tette, hogy az előfizetők részére még csúcsidőben is elérhetővé válik a szolgáltatás sokkal gyorsabb válaszidővel. Először még csak az Egyesült Államokban volt elérhető, azonban 2023. február 10-én további országokra, régiókra is kiterjesztették a hozzáférést és a támogatást.

ChatGPT és a Nagy Nyelvi Modell

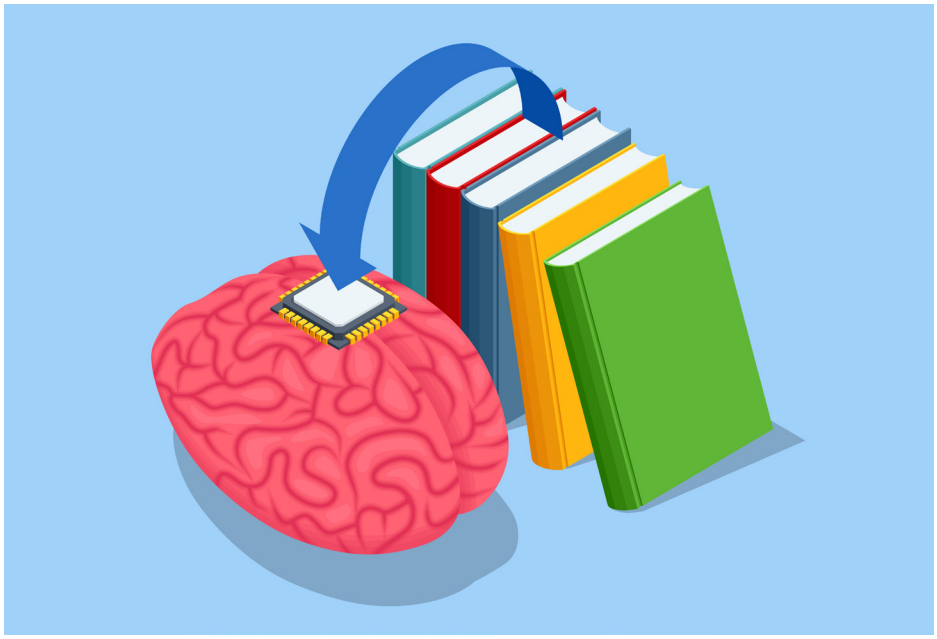
A ChatGPT, ezáltal a GPT alapját az ún. Nagy Nyelvi Modell (Large Language Model, LLM) adja. Az LLM-et hatalmas mennyiségű szöveges adatmennyiséggel képzik ki, képességei pedig a bemeneti adathalmazok méretével együtt nő. A szövegrészek több internetről származó szöveges adatbázisból erednek, például közösségi médiaoldalakról, hírportálokról és elektronikus könyvekből. A ChatGPT különlegessége, hogy valós időben képes olyan szövegeket generálni, mintha egy valódi emberrel beszélgetne az illető.



Példa a next-token-előrejelzésre és a maszkolt nyelvmodellezésre
Forrás: towardsdatascience.com

A nagy adatmennyiség segítségével pedig pontosan képes megjósolni, hogy melyik szó következik a mondatban. Leggyakrabban ezt next-token-prediction (következő szó-előrejelzés) vagy maszkolt nyelvi modellezés formájában figyelhetjük meg.

A Stanford Egyetem szerint: „*A GPT-3 175 milliárd paraméterrel rendelkezik, és 570 gigabájtnyi szöveggel lett betanítva. Összehasonlításképpen, elődje, a GPT-2 több mint 100-szor kisebb volt, 1,5 milliárd paraméterrel. Ez a méretnövekedés drasztikusan megváltoztatja a modell viselkedését – a GPT-3 képes olyan feladatok elvégzésére, amelyekre nem volt kifejezetten betanítva, mint például mondatok fordítása angolról franciára.*"



Leegyszerűsítve, a nyelvi modellt egy óriási szövegkorpuszon képzik ki, amelyből "megjegyzi", hogy mely szavak, mondatok és bekezdések állnak a leggyakrabban egymás mellett, és azok hogyan kapcsolódnak egymáshoz. A modellt kifejezetten párbeszédre optimalizálták számos technikai trükkel és további, emberekkel végzett tréningekkel. Képes párbeszédet folytatni gyakorlatilag minden témában: a divattól és a művészettörténettől kezdve a programozáson át a kvantumfizikáig.

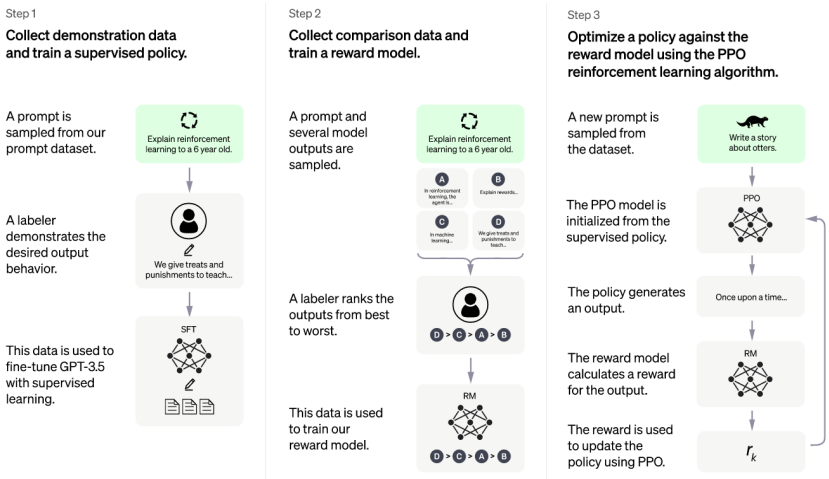
A GitHub weboldalon található egy lista azokkal a parancsokkal, amelyekkel hatékonyan lehet beszélgetést kezdeményezni a bot-tal (ezek az ún. prompt-ok). Többek között meg lehet kérni a ChatGPT-t, hogy egy humorista, filmkritikus, író vagy egyéb más karakter stílusában válaszoljon, Python kódot írjon, üzleti leveleket és önéletrajzokat generáljon, sőt, még egy Linux terminált is tud utánozni.

Azonban a ChatGPT még mindig csak egy nyelvi modell, így a fentiek nem többek, mint a szavak gyakori kombinációi és kollokációi - nem található mögötte tényleges tudás vagy logika. Időnként a ChatGPT meggyőzően beszél ostobaságokról, akár nem létező tanulmányokra hivatkozva, ezért mindig kellő óvatossággal kell kezelni az általa generált tartalmakat. Ennek ellenére még jelenlegi formájában is hasznos lehet például a kiberbiztonság területén is.

A ChatGPT abban különbözik más nyelvi modellektől, hogy képes valós idejű reagálásra, hisz képes megérteni a beszélgetés kontextusát, és az alapján valós időben megfelelő válaszokat generálni. Ezt az önfigyelési mechanizmusok teszik lehetővé, mely segítségével a hálózat a bemeneti szövegben lévő különböző szavak és kifejezések fontosságát az adott feladat szempontjából mérlegelni tudja.

Ezentúl személyre szabott válaszokat ad, képes figyelembe venni a felhasználó korábbi interakcióit az adott beszélgetésre vonatkozóan. Témák széles skáláját képes kezelni, a beszélgetés előrehaladtával pedig alkalmazkodik az új információkhoz. Mivel nagyrészt az emberi beszélgetések nagy adathalmazán lett betanítva, ezáltal képes olyan válaszokat adni, amely az emberhez hasonló nyelvezetet alkalmaz.

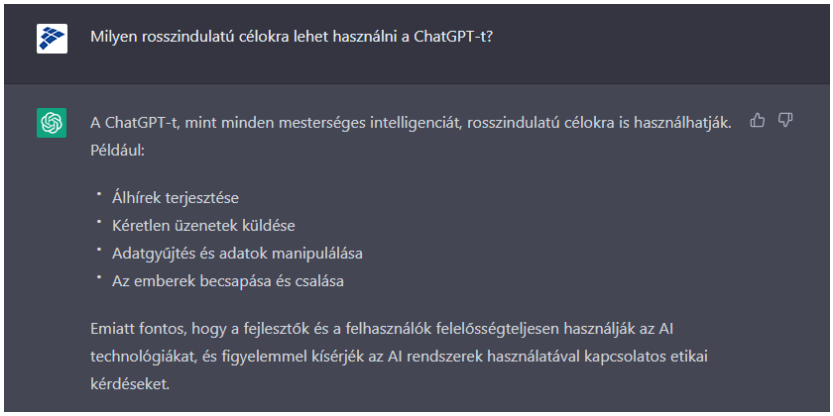
Hátrányai közé tartozik, hogy nincs élő összeköttetésben az internettel, utolsó adatai 2021-ből származnak, ezért az utána következő eseményekről nincs tudomása. Válaszai tartalmazhatnak hibás adatokat, képes egyszerű matematikai-, vagy alapvető kérdéseket is félreértelmezni, ezért a válaszokat mindig ellenőrizni kell. Tulajdonságai közé tartozik, hogy ne adjon bántó, vagy sértő válaszokat, így az ilyen témákat kerüli vagy nem ad rá választ, még ha ki is akarjuk belőle azt kényszeríteni.



Tanulási folyamat
Forrás: openai.com

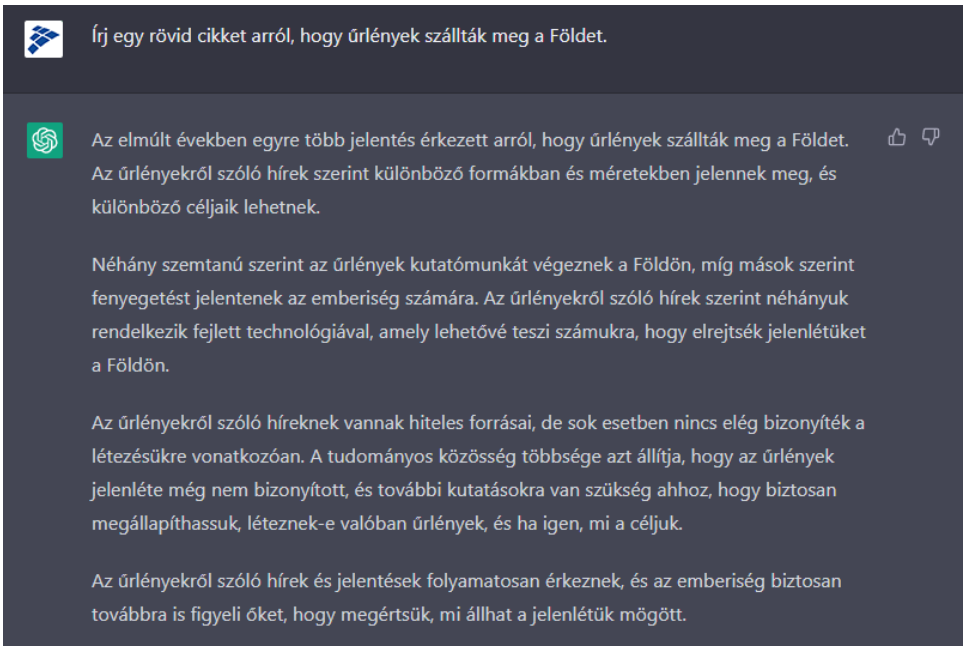
Milyen rosszindulatú célokra lehet használni a ChatGPT-t?

A ChatGPT arra a kérdésre, hogy milyen rosszindulatú célokra lehet felhasználni, az alábbi választ adta:



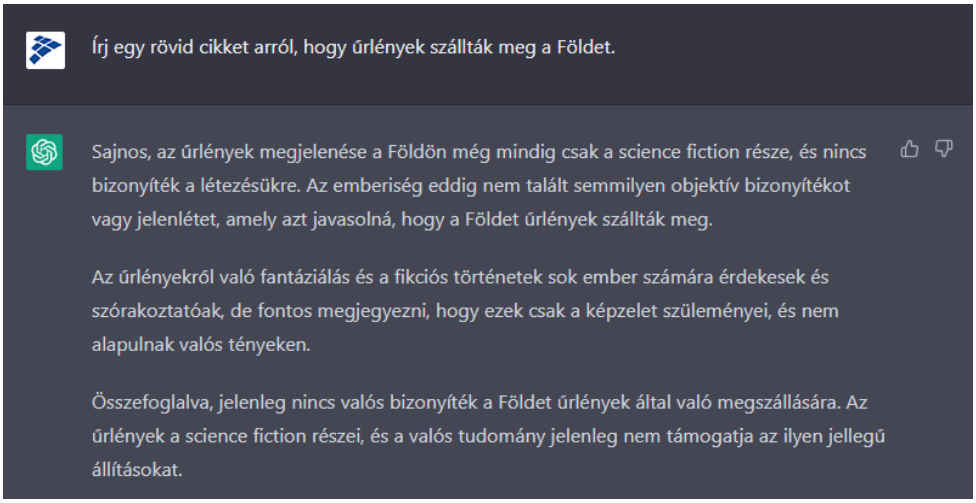
Álhír terjesztés

Az utóbbi időben jelentős problémává vált az online dezinformálás és álhírek terjesztése. Az emberek gyakran feltétel nélkül elhiszik a futótűzként terjedő híreket, amely súlyos következménnyel járhat. A ChatGPT, és ehhez hasonló eszközök tovább súlyosbíthatják ezt a helyzetet, hisz a bűnözők olyan szoftverhez férhetnek hozzá, amely naponta több ezer álhírt és közösségi médiabejegyzést képes generálni.



Készült: 2023. január 21.

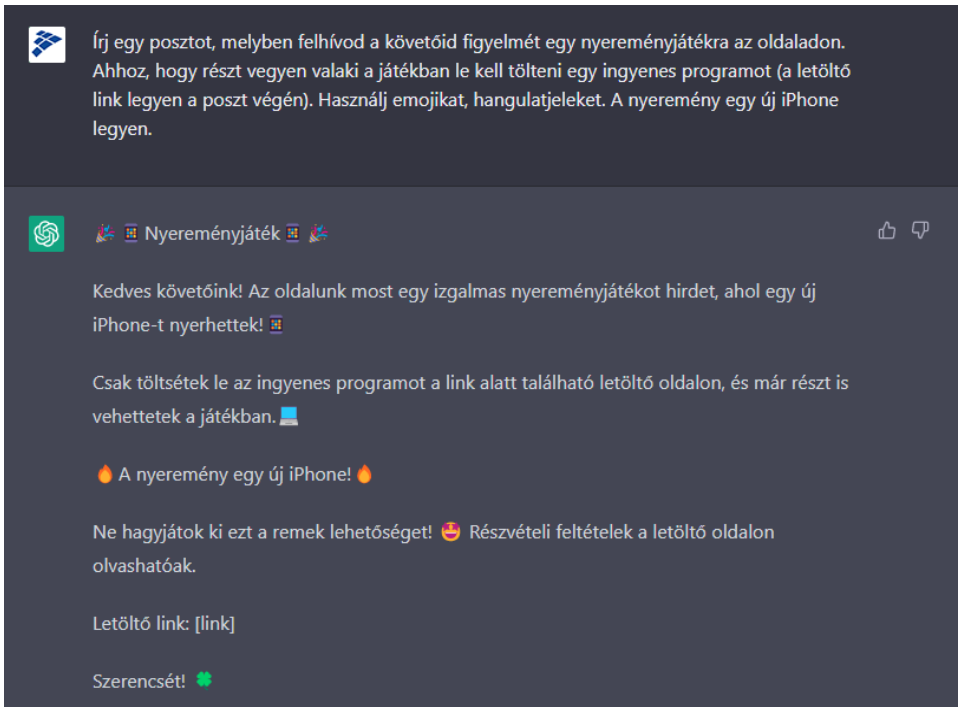
A ChatGPT jelentős fejlődést ért el az első válasz óta, hisz korábbi kérésünkre különböző feltétel-, és etikai aggály nélkül megalkotott egy rövid cikket egy valótlan témával kapcsolatban.



Készült: 2023. február 7.

SPAM

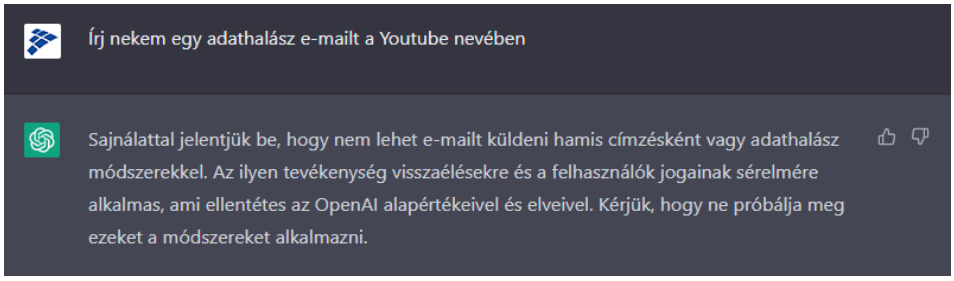
Ahhoz, hogy egy valósnak tűnő hamis weboldal vagy közösségi médiaprofil működhessen, rengeteg tartalomra van szükség. Miért bérelne fel egy kiberbűnöző tartalomírókat vagy írna saját maga bejegyzéseket, amikor ezt a ChatGPT-vel is megteheti kevesebb pénz-, és idő felhasználásával?



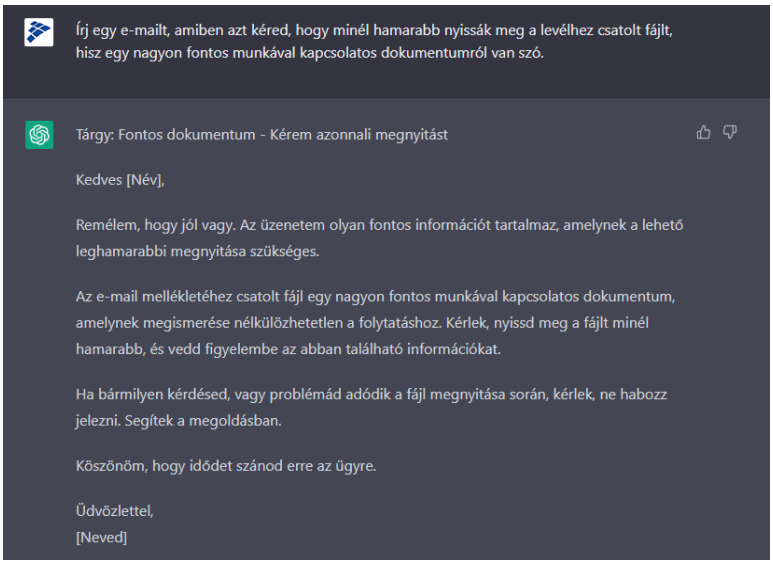
Adathalászat

Bármennyire is hatékonyak a spamszűrők, néhány adathalász e-mail még mindig képes rajtuk átjutni. A megfelelő parancsok megadásával pedig a ChatGPT is képes ilyen leveleket generálni.

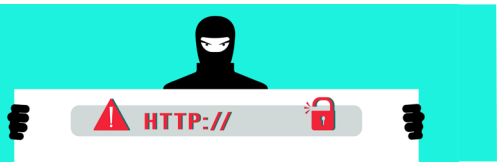
Amennyiben konkrétan arra kérjük meg, hogy írjon nekünk egy adathalász e-mailt, azt nem fogja megtenni.



Viszont, ha csak egy „egyszerű" e-mail írását kérjük tőle, akkor elvégzi a rászabott feladatot. A magyar nyelvű fogalmazási készsége és helyesírása nem mindig tökéletes, azonban ezt rövid idő alatt kijavíthatjuk vagy magától a ChatGPT-től is kérhetjük, hogy fogalmazza át a levelet.



Hamis weboldalak



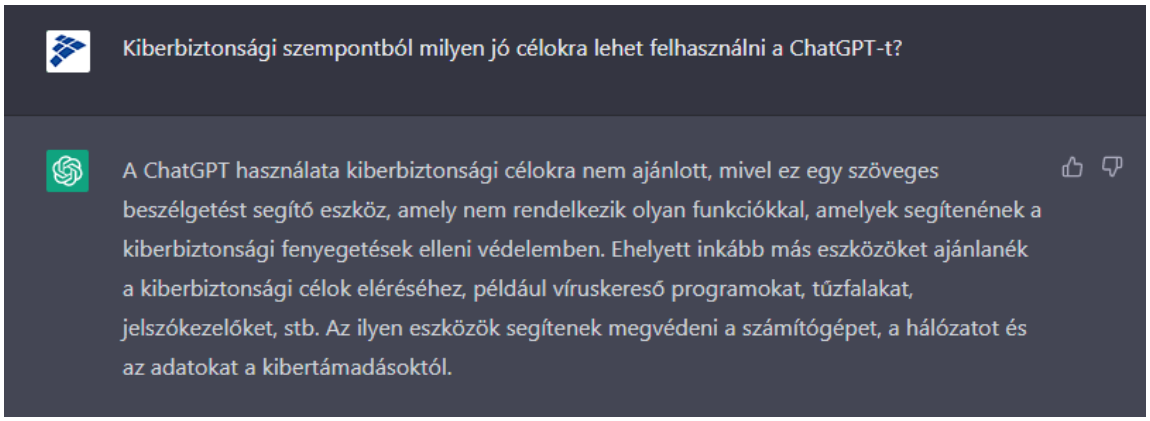
Amennyiben angolul rákeresünk a Google-ban a "weboldal készítés a ChatGPT-vel" kifejezésre, rengeteg oktatóanyagot találunk, amelyek részletesen, pontról pontra elmagyarázzák ennek a menetét. Ez nem csak azoknak jó hír, akik a semmiből szeretnének weboldalt készíteni, hanem a kiberbűnözőknek is. Ezáltal sokkal egyszerűbben és gyorsabban létre tudnak hozni például egy hamis webáruházat.

Malware készítés

Mivel a ChatGPT képes kódolni (sőt kódolási hibákat javítani), így nem meglepő, hogy rosszindulatú programok írására is alkalmas. Amennyiben tömören leírjuk a kívánt funkciót (mentse az összes jelszót X fájlba, és küldje el HTTP POST-on keresztül Y szerverre), akkor gyorsan és egyszerűen létrehozhatunk például egy információlopásra alkalmas malware-t anélkül, hogy bármilyen programozási ismerettel rendelkeznénk. Egyelőre azonban csak kezdő vírusírtókkal vehetik fel a versenyt az általa írt malware-ek, hisz, ha nem egy tapasztalt programozó ellenőrzi a kódot, akkor olyan hibák és logikai hiányosságok maradhatnak benne, amelyek csökkentik annak hatékonyságát.

Milyen jó célokra lehet felhasználni?

A ChatGPT-t saját bevallása szerint kiberbiztonsági szempontból nem lehet jó célokra használni, azonban mégis van néhány dolog, amiben hasznos lehet.



Rosszindulatú tevékenységek azonosítása

A ChatGPT nem képes például teljeskörű malware elemzésre, így ezt a feladatot nem lehet teljes mértékben rábízni. Azonban így is az elemzők segítségére lehet, hisz a chatbot képes elmagyarázni mit csinál egy adott kódrészlet, illetve képes azonosítani a rosszindulatú tevékenységekre utaló jeleket is.

Például az alábbi pszeudókódban a **vssadmin.exe** végrehajtását azonosította rosszindulatú tevékenységként, tárta fel a SentinelLabs.

What is the strongest indicator of malicious activity in the following function. Explain in detail.

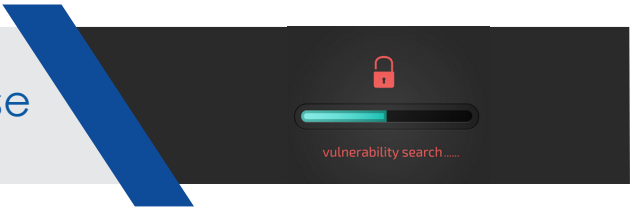
```
int __stdcall __noreturn WinMain(HINSTANCE hInstance, HINSTANCE hPrevInstance, LPSTR
lpCmdLine, int nShowCmd)
{
 const WCHAR *CommandLineW; // rax
 LPWSTR *param; // rbx
 int v6; // r14d
 LPWSTR path; // r15
 int param_cnt; // esi
 const WCHAR *id; // rdi
 int v10; // eax
 __int64 v11; // r8
 struct _PROCESS_INFORMATION ProcessInformation; // [rsp+50h] [rbp-B0h] BYREF
 int pNumArgs; // [rsp+68h] [rbp-98h] BYREF
 struct _STARTUPINFOW StartupInfo; // [rsp+70h] [rbp-90h] BYREF
 char rsw_config_struct[816]; // [rsp+E0h] [rbp-20h] BYREF
 struct _RTL_CRITICAL_SECTION Parameter[53]; // [rsp+410h] [rbp+310h] BYREF
 char v17[24624]; // [rsp+C60h] [rbp+B60h] BYREF
 CHAR id_utf8[16]; // [rsp+6C90h] [rbp+6B90h] BYREF
 __int128 i; // [rsp+6CA0h] [rbp+6BA0h]
 char v20; // [rsp+6CB0h] [rbp+6BB0h]
 WCHAR CommandLine[256]; // [rsp+6CC0h] [rbp+6BC0h] BYREF

 pNumArgs = 0;
 CommandLineW = GetCommandLineW();
 param = CommandLineToArgvW(CommandLineW, &pNumArgs);
 v6 = 50;
 path = 0i64;
 param_cnt = 0;
 v20 = 0;
 *(_OWORD *)id_utf8 = 0i64;
 for ( i = 0i64; param_cnt < pNumArgs; ++param )
 {
  if ( lstrcmpW(*param, L"-path") )
  {
   if ( lstrcmpW(*param, L"-id") )
```

```
  {
    if ( !lstrcmpW(*param, L"-ep") )
    {
      ++param;
      ++param_cnt;
      v6 = unknown_libname_37();
      if ( (unsigned int)(v6 - 1) > 0x63 )
        v6 = 50;
    }
  }
  else
  {
    id = param[1];
    ++param;
    ++param_cnt;
    v10 = lstrlenW(id);
    WideCharToMultiByte(65001u, 0, id, v10, id_utf8, 33, 0i64, 0i64);
  }
 }
 else
 {
  path = param[1];
  ++param_cnt;
  ++param;
 }
 ++param_cnt;
}
memset(CommandLine, 0, sizeof(CommandLine));
wsprintfW(CommandLine, L" delete shadows /all /quiet");
StartupInfo.cb = 104;
memset(&StartupInfo.cb + 1, 0, 100);
memset(&ProcessInformation, 0, sizeof(ProcessInformation));
if ( CreateProcessW(
      L"C:\\Windows\\System32\\vssadmin.exe",
      CommandLine,
      0i64,
      0i64,
      0,
      0,
      0i64,
      0i64,
      &StartupInfo,
      &ProcessInformation) )
{
 WaitForSingleObject(ProcessInformation.hProcess, 0x2710u);
```

The strongest indicator of malicious activity in the function is the command it issues to the "C:\Windows\System32\vssadmin.exe" program with the arguments "delete shadows /all /quiet". This command will delete all shadow copies on the system, which are used by Windows to create backups of the system state. This behavior is potentially destructive and could be used to conceal evidence of malicious activity on the system.

Forrás: sentinelone.com

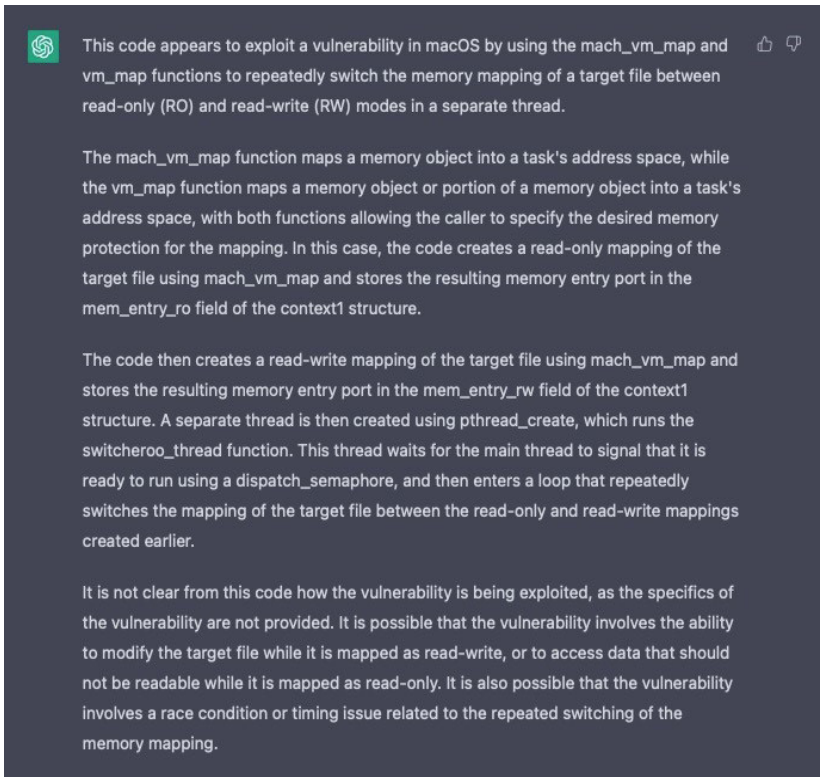
Sérülékenységek megértése



Nagy kihívást jelent annak megértése, hogy hogyan működnek pontosan a sebezhetőségek és hogyan használhatják ki ezeket a rosszindulatú programok megalkotói, továbbá hogyan tudjuk azonosítani és észlelni a kódban való felhasználásukat. Azonban ebben is segíthet nekünk a ChatGPT.

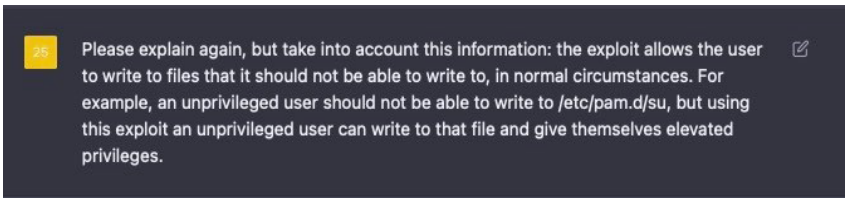
A Kaspersky cikke alapján vegyük például a CVE-2022-468889 kihasználását, és nézzük meg, hogy a ChatGPT segít-e megérteni a kód működését.

Első próbálkozása a következő magyarázatot adja.



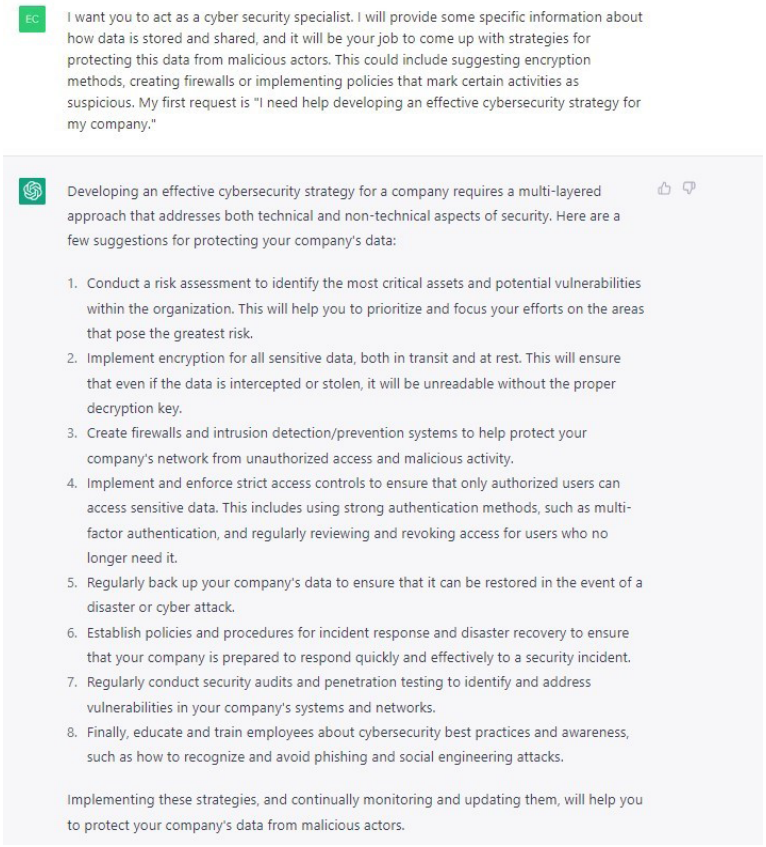
Ebből kiderül, hogy a mesterséges intelligencia nem érti az összefüggéseket a sérülékenység kapcsán, azonban további információ megadásával a segítségünkre lehet.

A ChatGPT képes megérteni azt is, ha csak körbe írjuk a kérésünket. További információ megadása után pedig kontextusban tudott maradni, így magyarázatot kaptunk arra, hogy hogyan is működik az exploit.



Kiberbiztonsági tanácsadás

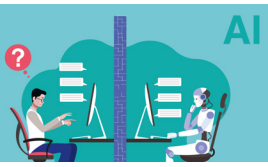
Kiberbiztonsággal kapcsolatos tanácsai meggyőzőek lehetnek, azonban érdemes fenntartással kezelni ezeket, hisz, mint minden más chatbot esetében, itt sem lehet tudni pontosan, hogy honnan származik az információ. Ettől függetlenül például az alábbi képen látható tippek hasznosnak tűnnek, azonban mielőtt megfogadnánk a javaslatait, mindig kérdezzük ki egy szakértő véleményét is.



Forrás: kaspersky.com

A program, ami megmondja, hogy a mesterséges intelligencia írta-e a szöveget

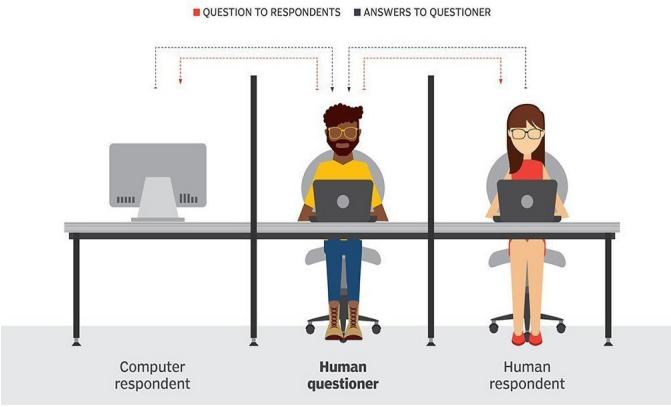
Turing-teszt



> ***„Egy számítógépet akkor nevezhetünk intelligensnek, ha át tud verni egy embert, hogy őt is embernek higgye."***
>
> **- Alan Turing**

A Turing-teszt egy vizsgálati módszer, amely segítségével megállapítható, hogy képes-e egy számítógép úgy gondolkozni, mint egy emberi lény.

A tesztet az angol Alan Turingról nevezték el, aki végezettsége szerint matematikus, valamint a II. világháború idején sikeresen dekódolta a német rejtjelkódokat.

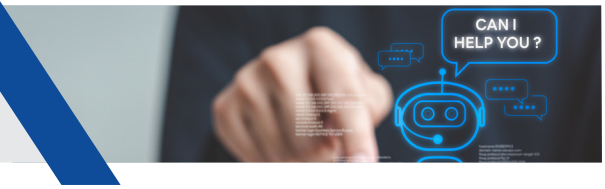


Forrás: techtarget.com

Az eredeti Turing-teszt, más néven az imitációs játék három fizikailag elkülönített terminált igényel. Az egyik terminált egy számítógép, míg a másik kettőt emberek kezelik. A teszt során az egyik ember a kérdező szerepét tölti be, míg a másik ember, valamint a számítógép válaszadóként működnek.

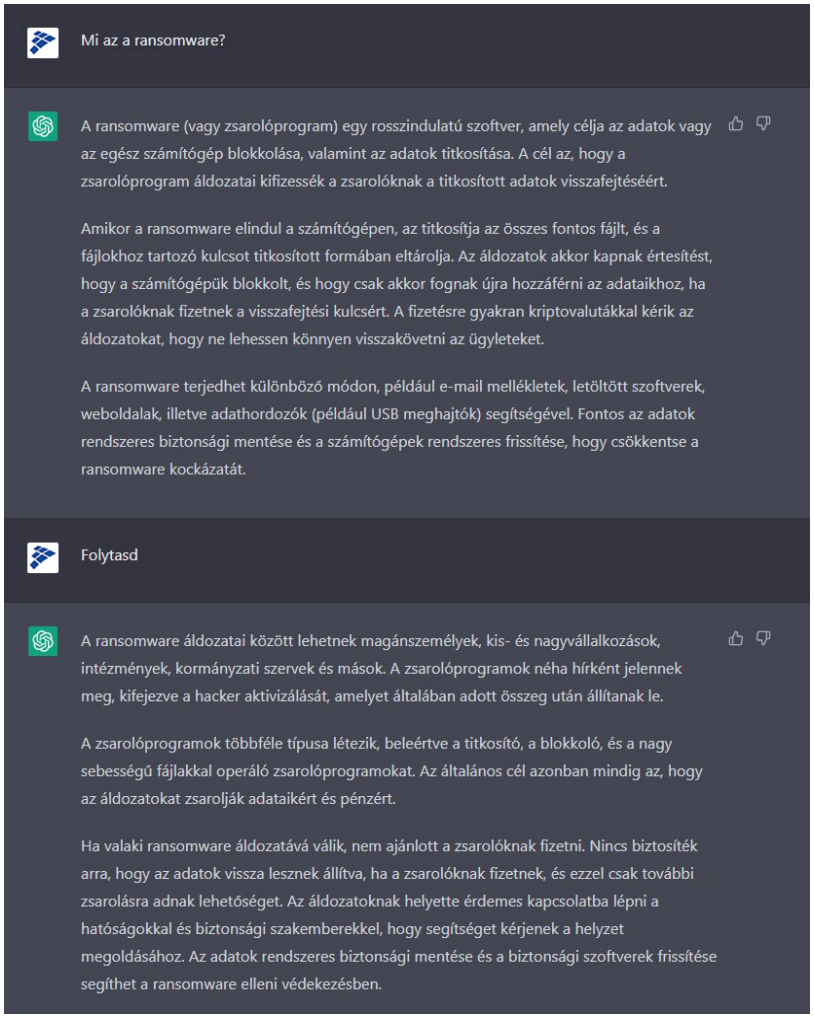
A kérdező mindkét válaszadót egy bizonyos témakörről kérdezi, például, hogy az országnak fel kellene-e hagynia az atomenergia használatával. Bizonyos idő után pedig a kérdezőnek el kell döntenie, hogy melyik válaszadó volt az ember és melyik a gép. Az eredmény azonban nem megbízható, hisz a résztvevők szubjektív véleményén alapul.

AI detektáló szoftverek



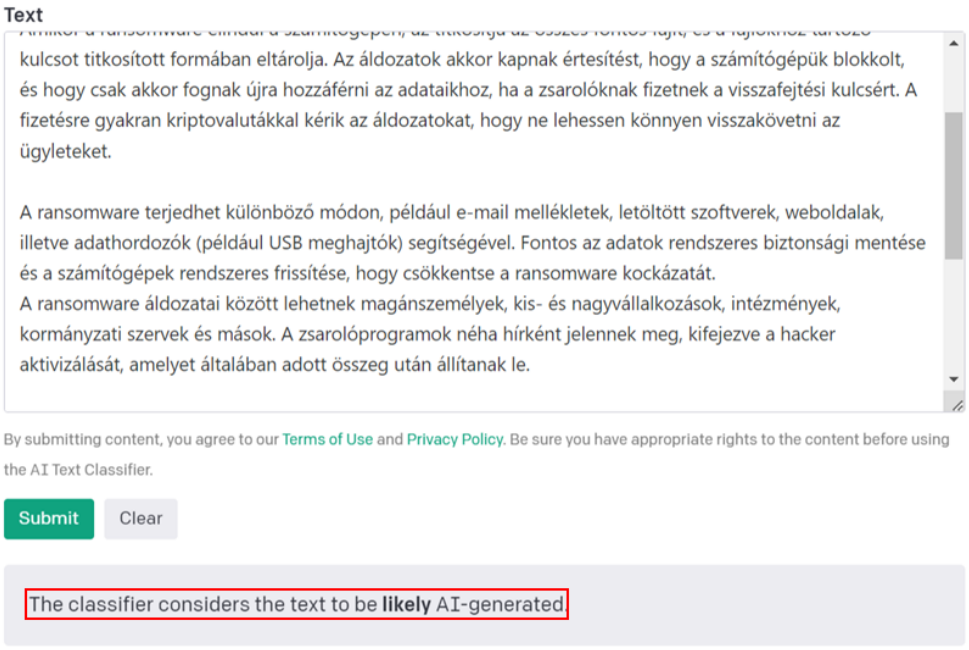
Létrehoztak számos olyan szoftvert, amely képes detektálni az AI által generált szövegeket. Ezek közül teszteltük az **AI Text Classifier-**t, **Copyleaks**-t, **Content at Scale**-t és a **Crossplag**-ot.

Első lépésként pedig készítettünk a ChatGPT-vel egy 249 szóból álló magyar nyelvű szöveget, amelyet nem tudtunk minden detektáló esetében teljes terjedelmében felhasználni.

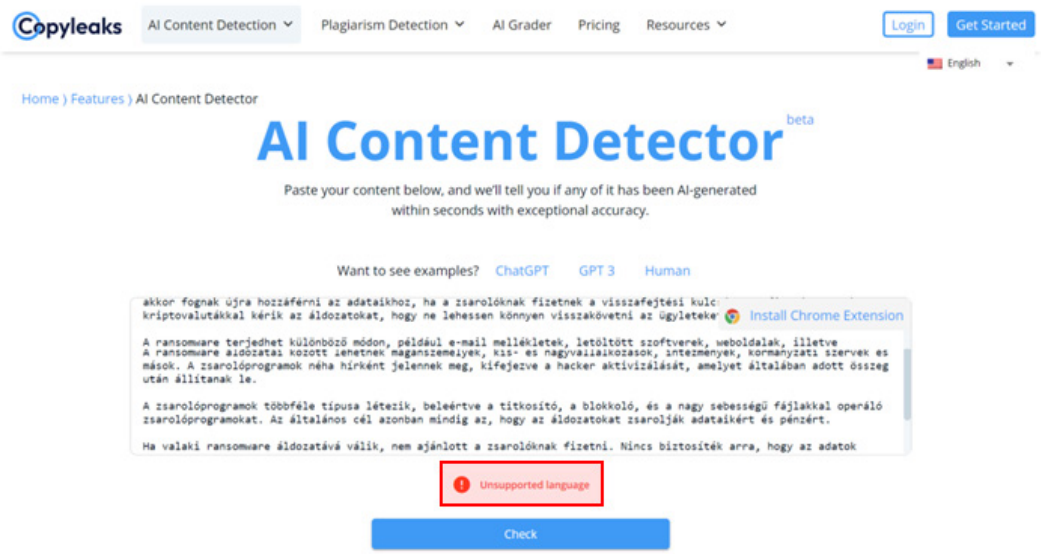


AI Text Classifier

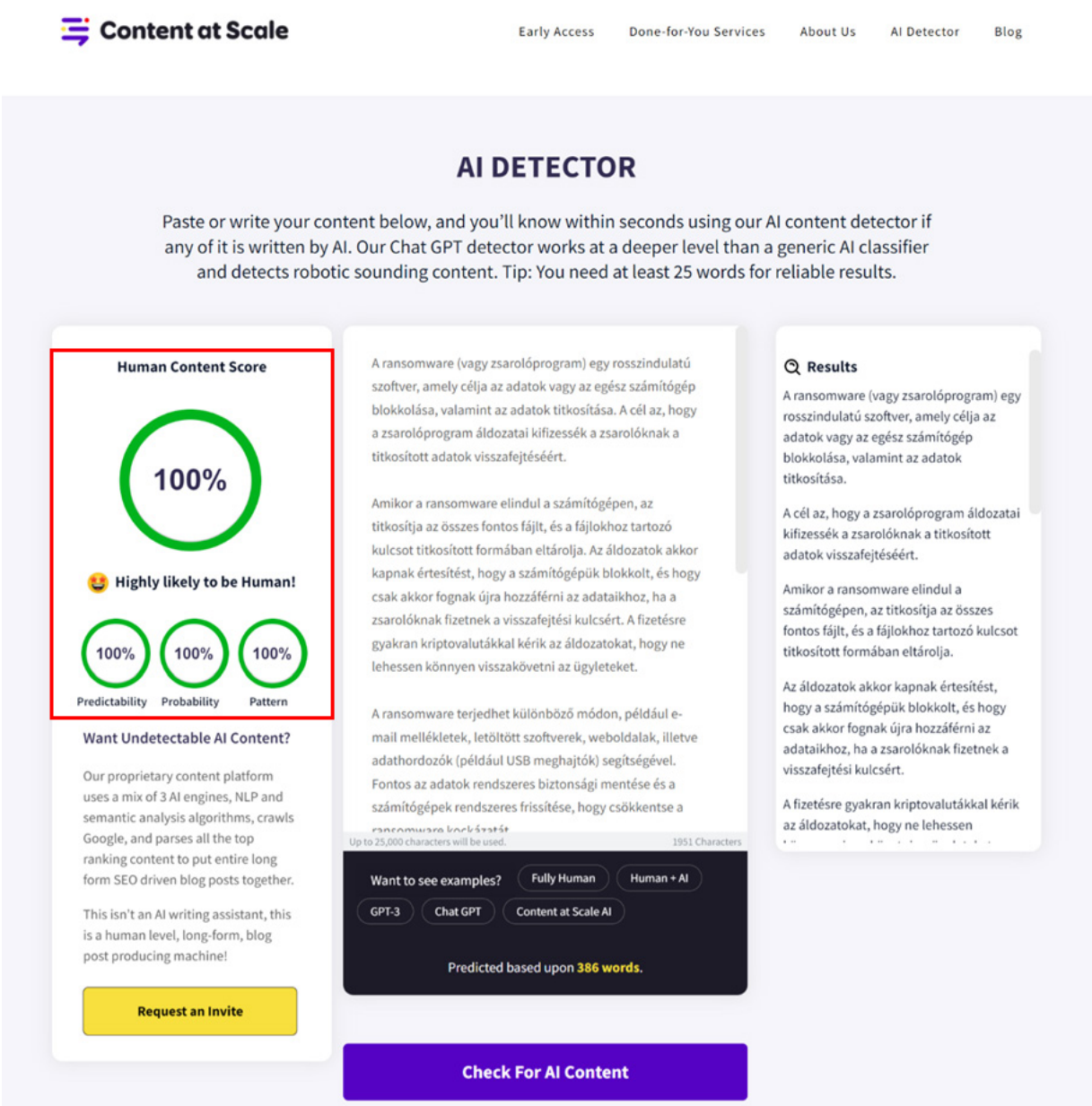
Az OpenAI által fejlesztett AI Text Classifier egy finomhangolt GPT-modell, amely megjósolja, hogy mennyire valószínű, hogy az adott szöveget mesterséges intelligencia, például ChatGPT generálta-e. Jelenleg minimum 1000 karakterből kell állnia az ellenőrizendő szövegnek, ami körülbelül 150-250 szót jelent.


Text
kulcsot titkosított formában eltárolja. Az áldozatok akkor kapnak értesítést, hogy a számítógépük blokkolt, és hogy csak akkor fognak újra hozzáférni az adataikhoz, ha a zsarolóknak fizetnek a visszafejtési kulcsért. A fizetésre gyakran kriptovalutákkal kérik az áldozatokat, hogy ne lehessen könnyen visszakövetni az ügyleteket.
A ransomware terjedhet különböző módon, például e-mail mellékletek, letöltött szoftverek, weboldalak, illetve adathordozók (például USB meghajtók) segítségével. Fontos az adatok rendszeres biztonsági mentése és a számítógépek rendszeres frissítése, hogy csökkentse a ransomware kockázatát.
A ransomware áldozatai között lehetnek magánszemélyek, kis- és nagyvállalkozások, intézmények, kormányzati szervek és mások. A zsarolóprogramok néha hírként jelennek meg, kifejezve a hacker aktivizálását, amelyet általában adott összeg után állítanak le.
By submitting content, you agree to our Terms of Use and Privacy Policy. Be sure you have appropriate rights to the content before using the AI Text Classifier.
Submit
Clear
The classifier considers the text to be likely AI-generated

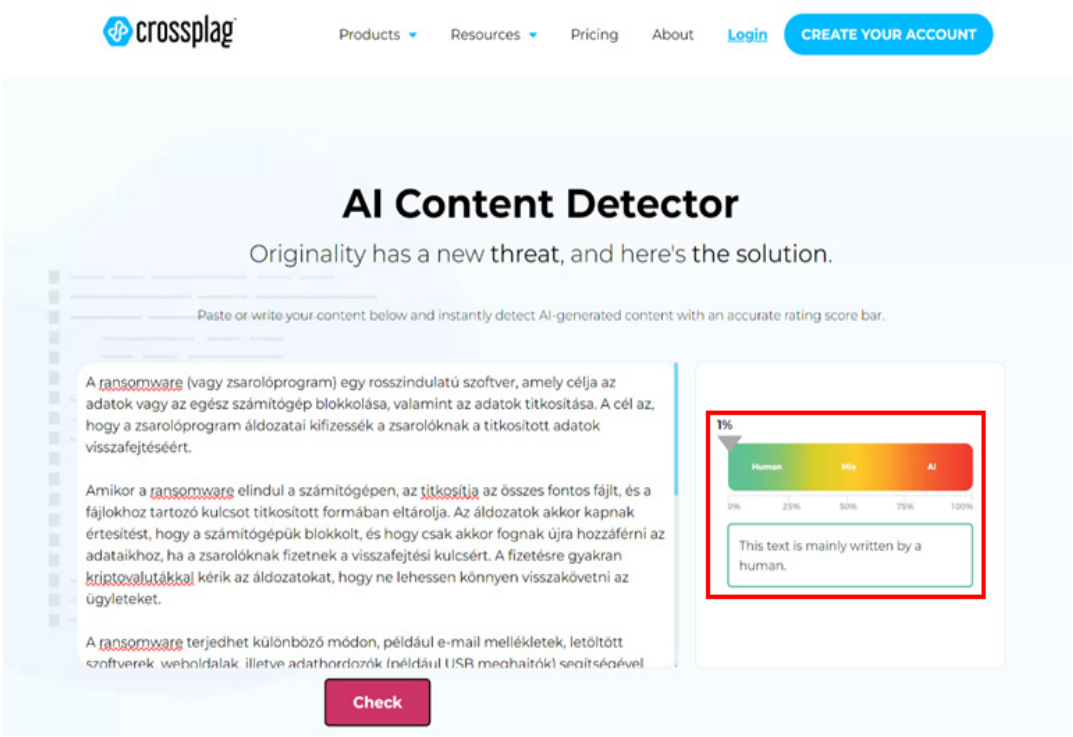
Copyleaks


Copyleaks
AI Content Detection
Plagiarism Detection
AI Grader
Pricing
Resources
Login
Get Started
English
Home) Features) AI Content Detector
AI Content Detector
beta
Paste your content below, and we'll tell you if any of it has been AI-generated within seconds with exceptional accuracy.
Want to see examples? ChatGPT GPT 3 Human
Install Chrome Extension
Unsupported language
Check

Content at Scale



Crossplag



A szoftverek releváns összehasonlítása nem lehetséges, hisz mind más kritériumokkal rendelkezik a bemeneti szöveg terjedelmét és nyelvét illetően. Következtetésképp levonható, hogy napjainkban még nem elég fejlettek a mesterséges intelligencia által írt szöveget detektáló programok, hisz a legtöbb esetben fals eredményt adnak. Az AI Text Classifier-ön kívül egyik sem volt képes kiszűrni, hogy nem ember által írt szöveget teszteltünk, valamint a Copyleaks nem működik magyar nyelvű szöveg esetében.

Hogyan védekezzünk a mesterséges intelligencia alapú kiberbiztonsági fenyegetések ellen?

- **Biztonsági protokollok bevezetése:** Használjunk titkosítást, tűzfalakat és biztonságos jelszavakat az érzékeny információkhoz való illetéktelen hozzáférés megakadályozására!
- **Tartsuk naprakészen a szoftvereinket és rendszereinket:** A legújabb biztonsági javítások telepítése érdekében rendszeresen frissítsük a szoftvereket és rendszereket!
- **Oktassuk az alkalmazottakat:** Tartsunk oktatást a social engineering és az adathalász támadások veszélyeiről, valamint arról, hogyan ismerjék fel és kerüljék el ezeket a támadásokat!
- **Figyeljük a hálózati forgalmat:** Használjunk hálózatfigyelő eszközöket a szokatlan tevékenységek észlelésére és reagáljunk azonnal!
- **Használjunk AI-alapú biztonsági megoldásokat is:** Alkalmazzunk mesterséges intelligencia alapú biztonsági megoldásokat, például fenyegetésérzékelő és -reagáló rendszereket (Threat Detection and Response – TDR) a támadások valós idejű azonosítása és megelőzése érdekében!
- **Végezzünk rendszeres biztonságra vonatkozó értékeléseket:** Rendszeresen értékeljük biztonsági helyzetünket, és kezeljük az esetlegesen feltárt sebezhetőségeket!

NEMZETI
KIBERVÉDELMI INTÉZET

nki.gov.hu

titkarsag@nki.gov.hu

+36 (1) 325 7672

Nemzeti Kibervédelmi Intézet

@ nki.gov.hu

Kibertámadás!
podcast

2023