

OTDK-dolgozat

2025

NEUMANN JÁNOS EGYETEM
MNB INTÉZET: FENNTARTHATÓ PÉNZÜGYEK KÖZPONT
NEMZETKÖZI GAZDASÁG ÉS GAZDÁLKODÁS MESTERSZAK



DETEKTÁLHATÓ-E AZ AI HASZNÁLATA A FELSŐOKTATÁSBAN?

CAN THE USE OF AI TOOLS BE DETECTED IN HIGHER EDUCATION?

Szabó Kornél László

Konzulens: Dr. Balogh Tamás László

Kézirat lezárásának dátuma: 2024.04.15.

DETEKTÁLHATÓ-E AZ AI HASZNÁLATA A FELSŐOKTATÁS-BAN?

ABSZTRAKT

A ChatGPT megjelenése után világszerte elterjedtek a különböző mesterséges intelligencia alapú eszközök. Az egyetemi világba is eljutottak, felmérések szerint a hallgatók nagyjából harmada használt már ilyen eszközt egyetemi feladatok megoldásához. Az AI modellek ilyen módon történő felhasználása csalásnak, akár plágiumnak is számíthat, emiatt megjelent az igény az AI-generált szövegek megbízható és hatékony detektálására.

Tanulmányomban hat darab ilyen célt szolgáló AI detektor szoftver hatékonyságát vizsgáltam meg 20 darab a GPT4 modell által generált, és 20 darab emberek által írt szövegrészleten keresztül. Összehasonlításképpen emberi értékelőkkel, egy egyetemi oktatóval és egy hallgatóval is ellenőriztettem a szövegeket. Ezen túlmenően pedig megvizsgáltam azt is, hogy könnyebben észlelhetőek-e a GPT4-nél régebbi nyelvi modellel generált szövegek.

Az ellenőrző szoftverek viszonylag hatékonynak bizonyultak, mindegyik legalább a szövegek 70 százalékának forrását eltalálta. Az AI-generált szövegeket majdnem minden esetben azonosítani tudták a GPT4 által generált szövegek esetén is, a GPT3.5 által generáltaknál pedig csak egy szoftver adott téves választ. Hibákba nagyobb arányban az emberek által írt szöveg észlelésekor futottak, ezek közül is többet AI-generáltnak minősítettek. Az emberi értékelők minden esetben rosszabbul teljesítettek a szoftveres megoldásoknál, elmondható, hogy szabad szemmel nem lehet pontos értékelést adni egy szöveg forrásáról.

A téves pozitív jelzések komoly problémát jelentenek, hiszen ezeket a detektor programokat sok esetben egyetemi oktatók használják a hallgatók munkájának ellenőrzésére. A téves pozitív jelzések hallgatók téves meggyanúsítását eredményezhetik, melynek következtében elbukhatják a tárgyat, vagy akár ennél komolyabb következményekkel is szembe nézhetnek.

TARTALOMJEGYZÉK

1.	BEVEZETÉS	1
1.1.	AZ AI FELISMERŐ SZOFTVEREK.....	2
1.2.	KORÁBBI KUTATÁSOK A TÉMÁBAN	4
2.	MÓDSZERTAN.....	7
2.1.	FELHASZNÁLT DETEKTOR SZOFTVEREK	9
2.2.	LIMITÁCIÓK	10
3.	EREDMÉNYEK	12
3.1.	ELLENŐRZŐ SZOFTVEREK ÉS EMBERI ÉRTÉKELŐK ÖSSZEHASONLÍTÁSA	12
3.2.	GPT3.5 ÉS GPT4 MODELLEK ÁLTAL GENERÁLT SZÖVEGEK ÖSSZEHASONLÍTÁSA	14
4.	ÖSSZEGZÉS.....	16
	IRODALOMJEGYZÉK	18

TÁBLÁZAT- ÉS ÁBRAJEGYZÉK

ábra 1: Az értékelőkhöz tartozó pontosság (saját szerkesztés).....	12
ábra 2: Az értékelőkhöz tartozó szenzitivitás (saját szerkesztés)	13
ábra 3: Az értékelőkhöz tartozó specificitás (saját szerkesztés).....	13
ábra 4: Szoftveres és emberi értékelők átlagos eredményének összehasonlítása (saját szerkesztés).....	14
ábra 5: Ellenőrző szoftverek pontossága a GPT4 és GPT3.5 modellek által generált szövegek értékelésekor (saját szerkesztés)	15

1. BEVEZETÉS

Az utóbbi egy évben elárasztották az internetet az olyan posztok, melyben az író arról számol be, hogy egy egyetemi kurzusra készült beadandó feladatával kapcsolatban csalással vádolja az oktató. A vádat egy AI detektor szoftver által adott eredményre alapozzák, mely értékeli, hogy egy beillesztett szöveg milyen valószínűséggel mesterséges intelligencia által generált. Az írók jellemzően határozottan állítják, hogy nem vették igénybe a mesterséges intelligencia segítségét, az ellenőrzőszoftver pedig tévesen adott pozitív eredményt. Ennek következtében ártatlanul kell súlyos következményekkel szembenéznük.

A mesterséges intelligencia (melyre tanulmányomban az elterjedtebb angol nyelvű rövidítést, az AI-t használok), azon belül is a nagy-nyelvi modellek témája 2022 végén tört be a köztudatba, miután az OpenAI november 30-án elérhetővé tette a ChatGPT nevű platformját. Ez egy a GPT nevű nagynyelvi modell 3.5-ös változatára épülő, felhasználóbarát chatbotot jelentett, mely az emberi nyelvet nagy pontossággal utánzó szöveges válaszokat ad a felhasználó kérdései és utasításai („prompt”-jai) szerint. Az oldal januárra már több, mint százmillió felhasználóval rendelkezett, amivel a történelem addigi leggyorsabban növekvő felhasználószámú szoftvere lett. (Hu, 2023)

Bár az ilyen rendszerek előzményei egészen a 20. század közepéig nyúlnak vissza (Foote, 2023), és a ChatGPT-t gyártó OpenAI is több modellt kiadott már korábban (GPT-1 - 2018, GPT-2 - 2019, GPT-3 – 2020), de az igazi áttöréshez a 2020. végén megjelent GPT3.5-ig kellett várni. Azóta pedig 2023. márciusában megjelent az újabb, még fejlettebb modell, a GPT4. (O’Neill, 2023)

Az AI modellek az oktatásban is felkapottá váltak. Oktatók számára is hasznos lehet a tananyag összeállításában, de talán még jelentősebb hatása volt a hallgatók életére, akiknek hatékony segítség lehet a tananyag megértésében, azonban a visszaélés lehetőségét is megteremti, azaz a tananyag elsajátítása helyett történő használatát. Felmérések szerint a hallgatók nagyjából harmada használt már valamilyen AI eszközt egyetemi feladatok megoldásához. (Henry, 2023)

Az AI által generált szövegek használata a plágium egy új formája. Bár nem egy másik szerző művét használja fel az író, de nem saját maga által írt szöveget állít be annak. (Crawford, 2023). Ennek észlelése azonban nagyobb kihívást jelent, mint létező műből történő

másolás esetén, hiszen a szoftver által generált szöveg nem érhető el nyilvánosan, és nem garantálható, hogy kérésre ugyanabban a formában generálja újra a tartalmat.

A használat felismeréséhez detektor szoftverek jelentek meg, melyek azt állítják, hogy képesek megkülönböztetni egymástól a mesterséges intelligencia által generált és az emberek által írt szövegeket. Készített ilyen a ChatGPT-t gyártó OpenAI, független szereplők, illetve a korábban plágiumdetektálásra használt szoftverek gyártói is építettek be rendszereikbe a mesterséges intelligencia érzékelésére alkalmas funkciót.

Pontosságukkal kapcsolatban az elmúlt évben több teszt is megjelent, illetve a fent írtak szerint rengeteg beszámolót tesznek közzé téves pozitív eredményekről. Természetesen nem tudhatjuk, hogy az ilyen posztok írói őszinték-e, azonban az esetek száma mindenképpen felveti a gyanút, hogy a szoftverek oldalán is probléma lehet. A témával mindenképpen érdemes foglalkozni, hiszen egy hamisan megvádolt hallgató nagyon súlyos következményekkel nézhet szembe, vagy legalábbis bonyolult procedúrán kell végigmennie, hogy ártatlanságát tisztázza. Ez komoly pszichológiai terhet jelenthet az érintett számára. (Gegg-Harrison, 2024)

A probléma miatt több AI detektor oldalán megjelenítenek figyelmeztetést, hogy a kapott értékelés önmagában nem alkalmas arra, hogy az alapján diákokat szankcionáljanak. A GPTZero tippeket is feltüntet a jobb ellenőrizhetőség és a hamis vádak elkerülése érdekében, például, hogy a hallgatók a Google Docs felületén írják a beadandó dolgozataikat, mert itt nyomon követhető a szövegek írásának időbeli haladása, ami meggyőző jele lehet, hogy a saját munkájukról van szó. (GPTZero, 2023)

Tanulmányomban 40 darab, fele-fele arányban AI-generált és emberek által írt szövegrészlet segítségével vizsgálom hat népszerű AI detektor szoftver hatékonyságát, illetve két emberi értékelő képességét a GPT által írt tartalmak azonosítására. A korábbi, a témában készült írásokhoz képest fontos különbség, hogy egy kutatásban, és azonos mintán vizsgálom a programok és az emberi értékelők hatékonyságát, ezáltal azok nagyobb biztonsággal hasonlíthatóak össze. Emellett megvizsgálom azt is, hogy mennyiben különbözik a szoftverek hatékonysága az aktuális GPT4, és az egyvel korábbi GPT3.5 modell használata esetén.

1.1. AZ AI FELISMERŐ SZOFTVEREK

A könnyen kezelhető, felhasználóbarát AI modellek megjelenésével és elterjedésével a nem-etikus felhasználásuk is a korábbiaknál nagyobb problémát jelent. Lényeges igény

alakult ki arra, hogy megkülönböztethetők legyenek az AI által generált tartalmak az emberek által készítettektől. Ezen igényre reagálva sorra jelentek meg olyan szoftverek, melyek a készítők állítása szerint nagy pontossággal képesek erre.

Készítettek ilyen független szereplők, más AI eszközök gyártásával foglalkozó cégek, plágiumellenőrző szoftverek gyártói, emellett egyes esetekben a nyelvi modellek fejlesztői tettek elérhetővé ellenőrző szoftvert. Például a ChatGPT-t üzemeltető OpenAI is készített ilyen, OpenAI AI Classifier néven. Azonban ezt 2023. nyarán leállították, mert a készítők saját tesztjeik alapján arra jutottak, hogy jelenleg nem lehet érdemben detektálni az AI használatát. (Kirchner, 2023)

Bár léteznek kevésbé elterjedt, alternatív módszerek is, a gépi tanulásra épülő, azaz a köznyelvben AI alapú ellenőrző szoftverek működési elvük szerint három fő csoportba tartoznak az OpenAI besorolása szerint. Ezek az egyszerű, „zero-shot”, vagy finomhangoláson alapuló osztályozók. Az egyszerű osztályozók esetén egy új modellt tanítanak be, hogy megkülönböztesse egy nyelvi modell válaszait a normál szövegektől. A „zero-shot” módszernél az adott nyelvi modellt használják, további betanítás nélkül, hogy észlelje a saját maga által írt szövegeket. A finomhangolásra alapuló módszer esetében pedig az adott nyelvi modellt további tréninggel fejlesztik a saját maga által írtak detektálására. (Solaiman, 2019)

Az értékelésben több szempontot is figyelembe vesznek az ellenőrző szoftverek, de a két legáltalánosabb és legelterjedtebb a „perplexity” és a „burstiness”. A perplexity azt fejezi ki, hogy a program által ismert nyelvi modellek alapján mennyire kiszámíthatóak a sorban egymást követő szavak, avagy mennyire meglepő ahhoz képest, amilyen szót ő helyezne oda. A burstiness pedig az ismétlődésekre utal, hogy milyen gyakran jelennek meg ugyanazok a minták a szövegben. (Xu, 2024)

A jelenleg elérhető detektor szoftverek több limitációval is rendelkeznek. Az OpenAI gyűjtése alapján elmondható (de más szoftverek gyártói is hasonló figyelmeztetéseket helyeztek el a honlapjaikon), hogy nem megbízhatóak az eredmények rövid, vagy az angoltól eltérő nyelven írott szövegek esetén, illetve kiszámítható (pl. számsorozatot, egyszerű programkódot tartalmazó), vagy a betanításhoz használt adatoktól jelentősen eltérő szövegek esetén. (Kirchner, 2023)

Problémát jelent, hogy létezik könnyen elérhető módszer az ellenőrző szoftverek kihasználására, a „paraphrasing attack”, azaz az úrafogalmazásos támadás módszere. Ennek lényege, hogy ha az AI-generált szövegeket átfogalmaztatjuk egy ezt a célt szolgáló algoritmusmal, a detektálás esélye jelentősen lecsökken. (Krishna, 2023)

Teoretikusan van rá lehetőség a „watermarking”-nak nevezett módszerrel, hogy a nyelvi modellek működése egy olyan algoritmus szerint történjen, melynek ismeretében egyértelműen beazonosíthatóak az általa létrehozott szövegek. Ez amellett, hogy a biztos észlelést lehetővé teszi, a téves pozitív jelzéseket is kiküszöböli. Azonban ennek feltétele, hogy a használt nyelvi modell az ellenőrző szoftver által ismert algoritmusra épüljön. Illetve a „paraphrasing attack” ezzel a módszerrel sem védhető ki. (Fu, 2023)

A detektor szoftverek a megjelenésük utáni első kutatásokban nem mutattak nagy pontosságot. (Pegorano, 2023). Emellett nagyobb visszhangot kapott a médiában néhány olyan eset, mikor valamely AI ellenőrző program AI-generáltnak nyilvánított egy-egy híres történelmi művet. Így tett az OpenAI ellenőrző szoftvere a Biblia egy részével (Chakravarti, 2023), míg a GPTZero az amerikai alkotmány elejét azonosította a mesterséges intelligencia által írottakkal. (Mitchell, 2023)

1.2. KORÁBBI KUTATÁSOK A TÉMÁBAN

Bár a téma meglehetősen friss, hamar felkeltette a kutatók figyelmét. Így az elmúlt másfél évben több kutatás is készült az AI használatának ellenőrizhetőségéről. Néhány fontosabb, a témában íródott tanulmányt mutatok be az alábbiakban.

Egy az International Journal for Educational Integrity-ben 2023. decemberében megjelent tanulmányban összesen 11, korábban a témával foglalkozó írást gyűjtöttek össze. Emellett ebben a tanulmányban is teszteltek detektor szoftvereket. Összesen 54 darab, pár-ezer karakteres szöveget készítettek, melyek öt kategóriára oszlottak: emberek által angolul írt; emberek által más nyelven írt, majd gépi fordítással lefordított; AI-generált; AI-generált és emberek által manuálisan szerkesztett; illetve AI-generált és AI által úrafogalmazott. Ezekkel 14 ellenőrző szoftvert teszteltek. Vizsgálatuk során az emberek által írt szövegeket érzékelték a legpontosabban (94%), a lefordított illetve AI-generált szövegek esetén már csak 65-70%-ban született helyes eredmény, a manuálisan szerkesztett, illetve AI által úrafogalmazott írásoknál pedig 30% és 15% volt a pontosság. (Weber-Wulff, 2023)

Már 2023. szeptemberében megjelent egy korábbi tanulmány ugyanitt. Ebben a Qatar University kutatói 15 darab GPT4 által generált, 15 darab GPT3.5 által írt, és 5 darab emberek által készített szöveg segítségével vizsgálták 5 darab detektor szoftver hatékonyságát. Az egyes platformok eredményei nagyon eltérőek voltak. Egy szoftver sem találta el az összes szöveg forrását. Amelyik jól teljesített az AI-generált szövegek felismerésében, az az emberi szövegek észlelésében követett el tévedéseket, amelyik utóbbiban szuperált, az pedig a generált szövegekből hitt többet ember által írtnak. Egyetlen program, a CrossPlag nem adott egy tévesen pozitív választ sem, azonban emellé nagyon alacsony volt a szenzitivitása is, azaz az AI-generált szövegeket sem detektálta. A vizsgálatban megállapították azt is, hogy a GPT3.5 által írt szövegeket általánosan nagyobb pontossággal észlelték, mint a GPT4 által írottakat. Azonban a CopyLeaks szoftver esetén ez fordítva működött, előbbi kategóriából egyet sem, utóbbiból majdnem az összeset helyesen értékelte. (Elkhatat, 2023)

A Technical University of Darmstadt 2023. áprilisában 20 darab AI felismerő szoftvert és eljárást vetettek nagymintás vizsgálat alá. Összesen 58.546 ember által írt, és 72.966 GPT3.5 által generált szöveget gyűjtöttek különböző témákban, majd egy nyelvi modell segítségével kiválasztották a két csoport egymástól szerkezetükben leginkább különböző 10%-át. Ezeken keresztül vizsgálták a programok hatékonyságát. A vizsgálatban minden ellenőrző platform és eljárás az esetek kevesebb, mint felében tudta azonosítani az AI által írt szövegeket, míg az emberi írást jellemzően az esetek legalább 90%-ában sikerült. (Pegorano, 2023)

Arról is készültek kutatások, hogy emberi szemmel észrevehető-e az AI felhasználása. 2023. decemberében a Research Methods in Applied Linguistics folyóiratban jelent meg egy erről szóló tanulmány. Ebben 72 értékelő vett részt, akik mind a top 30 alkalmazott nyelvészeti témájú folyóiratnál cikkbírálok. Mindenki véletlen kiválasztással kapott négy absztraktot, melyek vagy AI-generáltak vagy ember által írtak voltak. Egy értékelő sem találta el az összes szöveg forrását. 13-an a négyből három szöveg esetén adtak helyes választ, a többiek, azaz az értékelők több, mint 80%-a az írások felét, vagy annál is kevesebbet azonosított helyesen. 9 fő egyetlen helyes választ sem adott. (Casal, 2023)

Összességében elmondható, hogy a korábbi kutatásokban jellemzően nem tudták a szoftverek hatékonyan észlelni az AI-generált tartalmakat, míg az emberek által írt szövegeket pontosabban, de még mindig sok hibával tudták azonosítani. 2023 második fele óta már jelentősen jobban teljesítenek a generált szövegek észlelésében, azonban az emberek által írtak azonosításában nem javult tovább az eredményük. Míg a korábbi tanulmányok fő

megállapítása az ellenőrző programok általános pontatlansága, addig a későbbi művekben már a téves pozitív jelzések még mindig magas arányát emelik ki problémaként.

2. MÓDSZERTAN

Az első kutatási kérdésem, hogy az AI detektorok megbízhatóan mutatják-e ki, hogy egy szöveg AI-generált-e. Második kutatási kérdésem, hogy az AI detektorok pontosabban mutatják-e ki, hogy egy szöveg AI-generált-e, mint ahogyan emberek ránézésre észlelik.

Harmadik kutatási kérdésem, hogy a ChatGPT3.5 válaszait sikerebben detektálják-e a szoftverek, mint a 4-es verzió válaszait. A GPT3.5 által generált szövegeket az emberi értékelőkkel nem véleményeztettem, csak szoftveres teszt történik.

A fő vizsgálat mintáját összesen 40 darab gazdaságtudományi témájú szöveg jelenti, melynek fele emberek által írt, másik fele pedig a GPT4 modell által generált. Minden szöveg angol nyelvű, mivel az ellenőrző szoftverekről ismert, hogy eltérő nyelvű szövegekkel nem megbízhatóak az eredményeik (Kirchner, 2023). A szövegek hossza 239 és 447 karakter közt alakul (szóközökkel), a detektorszoftverek követelményeihez igazítva. Átlagos hosszuk 332,8 karakter.

Az emberek által írt szövegek kiválasztásakor szempont volt, hogy ha kizárni nem is lehet, minél alacsonyabb legyen a kockázata annak, hogy AI által generáltak, vagy az AI modell betanításakor használt adatok része voltak. Emellett szempont volt az is, hogy ne legyenek túlzottan szakmaiak, mert az AI modellek úgy tartják, hogy alapszakos-mesterszakos hallgató szintjén vannak, illetve nagyon technikai szövegeknél felismerhető, hogy ember által írtak. Ehhez az archxiv felületen, még nem megjelent tanulmányok közt kerestem 2022. októberben közzétett írásokat. Ez kellően közel van a GPT megjelenéséhez, hogy a betanításhoz használt adatokba már nem kerülhettek be, de akkor még nem jelent meg a ChatGPT, így azt a készítőik nem használhatták művük elkészítéséhez.

A modellnek a következő prompt-ot adtam angolul: „Írj egy bevezetőt egy gazdasági témájú tanulmányhoz, ebben a témában: (...). Tíz bekezdésnyi szövegnek kellene lennie, úgy, hogy minden bekezdés 300 és 400 karakter közti hosszal rendelkezzen (szóközöket is számolva). Írj egy mesterszakos hallgató stílusában.”. Az (...) helyére két téma került, melyek két általam alkalmasnak talált tanulmány címei voltak. Az egyik a „Macroeconomic evaluation of the growth of the UK economy over the period 2000 to 2019”, azaz az Egyesült Királyság gazdasági növekedésének makroökonómiai értékelése a 2000 és 2019 közti időszakban. A másik cím pedig az „Impact of Business Analytics and Decision Support Systems on e-commerce in Small and Medium Enterprises”, azaz az üzleti elemző és döntéstámogató rendszerek hatása az e-kereskedelemre a kis- és középvállalkozások körében.

Csak szoftveres tesztelésre készült 20 db szöveg a GPT3.5 modellel is, ugyanazzal az utasítással, mint a GPT4 esetében. Ezeken keresztül azt vizsgálom, hogy ezeket sikeresebben észlelik-e a detektor szoftverek a 4-es verzió által generált szövegeknél. Ezen szövegrészletek hossza 220 és 487 karakter között alakul, 344,05 karakter átlagos hosszal.

A széles intervallum mögött több ok is áll. Egyrészt egyik szoftver sem tartotta be a „prompt”-ban szereplő terjedelmi kérést, érkezett a megadott tartománytól alulról és felülről is eltérő hosszú szöveg. A GPT3.5-nál különösen jelen volt ez a probléma, illetve mellette még többszöri kérésre sem tudta értelmezni a szóközzökkel együtt számolt karakterszámot,. Emellett esetében a kért tíz bekezdés helyett is hét, illetve kilenc darabot generált, melyek így manuális szétszedésre szorultak. Ebből adódik a minta másik két csoportjához képest nagyobb szórás a szövegek hosszában.

A másik ok, hogy egy ember által írt vagy annak ható szöveg esetében természetes, hogy van változatosság a mondatok hosszát és számát tekintve. Emiatt szerettem volna a lehető legkevesebb módosítást eszközölni, nem akartam például félbevágni egy hosszabb mondatot, amiért a szöveg hossza kilóg miatta az elvárt tartományból.

Ennek ellenére a GPT3.5 által generált szövegekben két részre bontottam néhány bekezdést (a mondatokat nem módosítva), az emberek által írt szövegekből pedig eltávolítottam a hivatkozásokat, ábrákra történő utalásokat, illetve felsorolásokat. Ennek oka, hogy ezeket saját bevallásuk szerint nem kezeli jól a detektor szoftverek, illetve az emberi értékelőknek is valószínűleg gyanúsak lettek volna ezen részek.

Három ellenőrző szoftvernél a terjedelmi követelmények nem voltak nyíltan elérhetőek, így a szöveggenerálás után derült csak ki, hogy a minimum várt karakterszám magasabb, mint a legtöbb szöveg hossza a mintában. Ebben a CopyLeaks, a Winston.ai és a ZeroGPT szoftverek voltak érintettek. Így ezek esetében párosával kerültek tesztelésre a bekezdések, az eredmény pedig mindkét szövegrészlethez beírásra került. Azaz kétszer olyan hosszú szövegek alapján, és feleannyi alkalommal teszteltek. Ez felvet problémákat, hiszen a hosszabb szövegek pontosabb elemezhetőségéből adódóan előnnyel indulnak a többi programhoz képest, viszont minden tévedésük duplán számít.

Az értékelést hat darab detektor szoftver, emberi oldalról pedig egy hallgató, és egy oktató végezték. A szoftverek kiválasztásában azok népszerűsége és hírneve volt a fő szempont. A szövegeket véletlengenerált sorrendben kapták meg az értékelők, és nem kaptak információt a generált és emberek által írt szövegrészletek arányáról.

A szoftverek eltérően értékelik a szövegeket, van amelyik bináris választ ad (AI által generált avagy nem), van amelyik ötosztályú skálán értékeli (az AI generáltság valószínűsége szerint), illetve van amelyik százalékos eredményt ad (szintén a valószínűségére). A kapott válaszokat a bináris változatra standardizáltam, mely esetben az ötosztályú eredményből a felső kettő, azaz a valószínű és nagyon valószínű válaszok (60% feletti érték), százalékos eredménynél pedig a 65% feletti válasz jelenti, hogy a szöveget AI által generáltnak tartja az adott szoftver. Az emberi értékelők kétosztatú, azaz AI vagy emberi értékelést adtak az egyes szövegekre.

A kapott értékelések négy csoportra oszlanak, valós pozitív (AI-generáltnak nyilvánított ténylegesen AI-generált szöveg), tévesen pozitív (AI-generáltnak nyilvánított ember által írt szöveg), valós negatív (ember által írtnak nyilvánított ember által írt szöveg) és tévesen negatív (ember által írtnak nyilvánított AI-generált szöveg). Ezek alapján pontosság, specificitás és szenzitivitás eredményeket számoltam. A pontosság a helyes válaszok aránya az összes válaszhoz viszonyítva, a specificitás a valós pozitív válaszok osztva a valós pozitív és tévesen negatív válaszok összegével (azaz az összes AI generált szöveg számával), a szenzitivitás pedig a valós negatív válaszok osztva a valós negatív és tévesen pozitív válaszok összegével (azaz az összes emberek által írt szöveg számával).

Az így kapott eredmények alapján összehasonlításra kerül a szoftverek hatékonysága egymással, illetve az emberi értékelőkével.

2.1. FELHASZNÁLT DETEKTOR SZOFTVEREK

A vizsgálat során hat darab detektor szoftvert használok, ezek a Content at Scale, a CopyLeaks, a GPTZero, a Sapling, a Winston.ai és a ZeroGPT rendszerei. Kiválasztásukban elsődlegesen népszerűségük játszott szerepet. A programok mindegyike ingyenesen elérhető, bár közülük több csak korlátozott számú ellenőrzést enged elvégezni előfizetés nélkül.

A Content at Scale 2022. szeptembere óta biztosít AI detektor szolgáltatást, elsősorban piaci szereplőket és tartalomgyártókat célozva. Rendszerük háromféle válasszal (igen/nem/talán) értékeli, hogy a tartalmak AI-generáltak-e. Saját méréseik szerint ezt 98,3%-os pontossággal teszi. Emellett egy AI-generált szövegek átfogalmazására alkalmas szoftvert, illetve egy képeket vizsgáló AI detektort is működtetnek. (Content at Scale, 2022)

A CopyLeaks 2015 óta foglalkozik plágium- és AI detektálással. Programjuk kétosztatú skálán értékeli a szövegeket, saját méréseik szerint angol nyelven 99,86%-os pontossággal. Állításuk szerint kellően nagy mintán 0,2% a téves pozitív jelzés esélye, ami a

legkisebb arány az ismert platformok közül. Ennek okaként azt tüntetik fel, hogy felületük nem az AI-generált szövegekre, hanem az emberi nyelvhasználatra jellemző mintákat keresi. (CopyLeaks, 2023a). A szoftver prémium verziója 29 más nyelven is képes szövegek elemzésére, minden esetben 95% feletti pontosságot biztosítva. (CopyLeaks, 2023b)

A GPTZero 2023. januárja óta működik és a cég szerint a legnépszerűbb elérhető detektor szoftver a maga 2,5 millió felhasználójával. A rendszer annak valószínűségét adja meg százalékos formában, hogy a kapott szöveg AI-generált. Emellett háromelemű skálán jelöli azt is, hogy mennyire magabiztos az adott válaszban. (GPTZero, 2023)

A Sapling nevű cég elsődlegesen AI asszisztens szoftvereket gyárt, de emellett elérhetővé tettek honlapjukon egy AI ellenőrző platformot is. Százalékos formában adja meg annak a valószínűségét, hogy a szöveget a mesterséges intelligencia írta. Honlapjukon kiemelik, hogy semmilyen szoftver nem alkalmas arra, hogy önmagában biztosan megállapítsa, hogy az adott tartalom AI-generált-e. (Sapling, 2023)

A magát a legmegbízhatóbb programként hirdető Winston.ai honlapján 99.8%-os pontosságot ígér. A legtöbb szoftverhez képest fordítva működik, százalékos formában annak az esélyét adja meg, hogy a szöveget ember írta. Képeken, illetve szkennelt dokumentumokon szereplő szöveget is tud elemezni. (Winston.ai, 2023)

A ZeroGPT az általa AI-generáltnak gyanított részek százalékos arányát adja vissza egy szövegből. Ők maguk 98%-os pontosságot mértek, de folyamatosan fejlesztik a rendszert, és szeretnék 2% alá szorítani a hibás válaszok arányát. A honlapon feltüntetett információk szerint minden nyelven képes szövegek elemzésére, arról azonban nem számolnak be, hogy a pontosság eltér-e az angol szövegek tesztelésén keresztül kapott értéktől. (ZeroGPT, 2023)

2.2. LIMITÁCIÓK

A kutatásom rendelkezik korlátokkal, melyek negatívan befolyásolhatják az eredményt. Egyrészt, a vizsgálat alapjául szolgáló mintát mindössze 40 darab szöveg jelenti (illetve további 20 darab a GPT4 és GPT3.5 észlelhetőségének összehasonlításához). Az ellenőrző szoftvereket gyártó cégek által publikált pontossági arányt vélhetően nagyobb mintán történő tesztelés eredményeként kapták.

Másrészt a vizsgált szövegek mindössze egy bekezdés hosszúak, gyakran éppen csak elérve a rendszerek által igényelt minimális karakterek vagy szavak számát. A detektor

szoftverek pontossága azonban saját bevallásuk szerint a vizsgált szövegek hosszának növelésével javul. (Kirchner, 2023). Egyes szoftverek azért nem kerültek be a vizsgálatba, mert nem is használhatóak ilyen rövid szövegekkel, például az egyetemi világban elterjedt Turnitin rendszere minimum 300 szavas szövegeket fogad csak be. (Turnitin, 2023)

Három szoftver esetében a terjedelmi követelmény nem volt nyíltan elérhető, és a szövegek legenerálása után derült csak ki. Így ezek a már legenerált szövegekkel történő tesztelésükhöz kettesével csoportosítva kapták meg azokat. Ez a nagyobb terjedelemből adódó könnyebb elemezhetőség, illetve az esetleges hibák duplázódása miatt befolyásolhatta az adott programok által elért eredményt.

Harmadrészt a felhasznált emberi szövegek egyszerűbbek, tanulmányok bevezetőiből vannak. Előfordulhat, hogy szakmaibb, elemzést illetve eredményeket tartalmazó szövegrészek esetén könnyebben megkülönböztethetőek lettek volna az AI-generált bekezdésektől.

Végezetül, a vizsgálat során csak teljesen ingyenes, vagy ingyenes verziót is kínáló detektor szoftverek kerültek tesztelésre. Lehetséges, hogy más, csak fizetős szolgáltatást biztosító platformok, vagy a tanulmányomban vizsgált eszközök fizetős változatai másképpen működnek és nagyobb pontossággal bírnak.

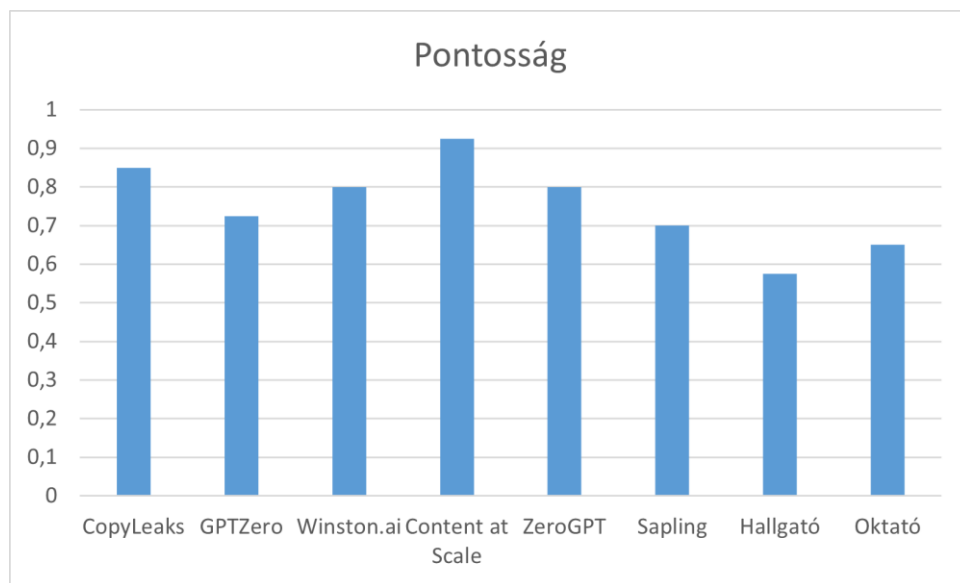
3. EREDMÉNYEK

3.1. ELLENŐRZŐ SZOFTVEREK ÉS EMBERI ÉRTÉKELŐK ÖSSZEHASONLÍTÁSA

A módszertani részben ismertetett hat AI detektor szoftver tesztelése 20 darab AI-generált és 20 darab emberek által írt szövegrészlet felhasználásával történt.

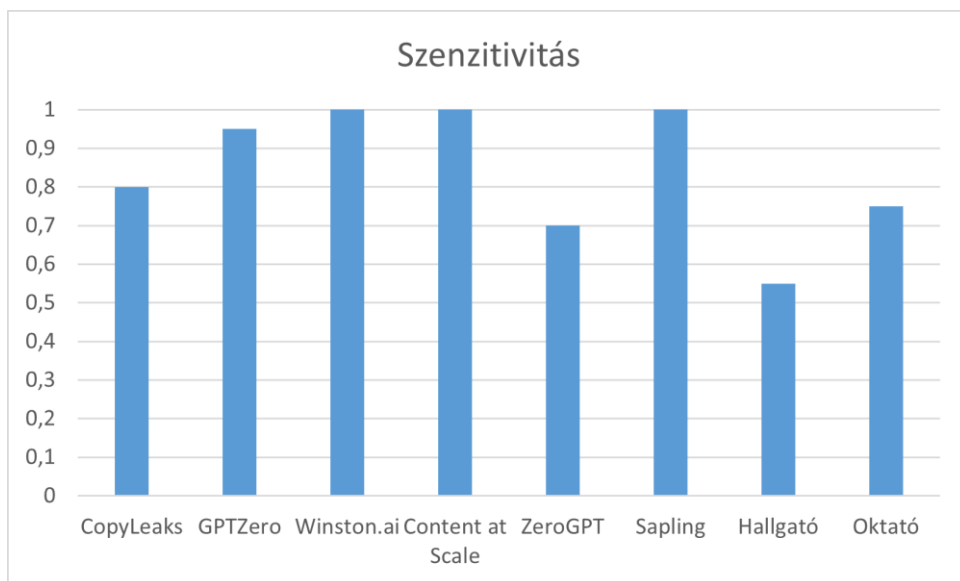
Az egyes ellenőrző programok válasza alapján számolt pontossági értékek láthatóak az 1. ábrán, mely megmutatja, hogy az összes válasz közül mennyi volt a helyesek aránya. Mindkét emberi értékelő, az oktató és a hallgató is rosszabbul teljesített az ellenőrző szoftverek mindegyikénél. Az oktató 65 százalékos, míg a hallgató 57,5 százalékos eredményt ért el.

Az összes szoftver esetében elmondható, hogy legalább a szövegek 70 százalékanak eltalálták a forrását. A legjobban a Content at Scale rendszere teljesített, 92,5 százalékos eredménnyel, ami azt jelenti, hogy mindössze három szövegrészlet esetében értékelt tévesen. A CopyLeaks 85, a Winston.ai és a ZeroGPT pedig egyaránt 80 százalékot ért el. A ZeroGPT és a Sapling a mezőnytől lemaradva 72,5 és 70 százalékos eredménnyel végzett.



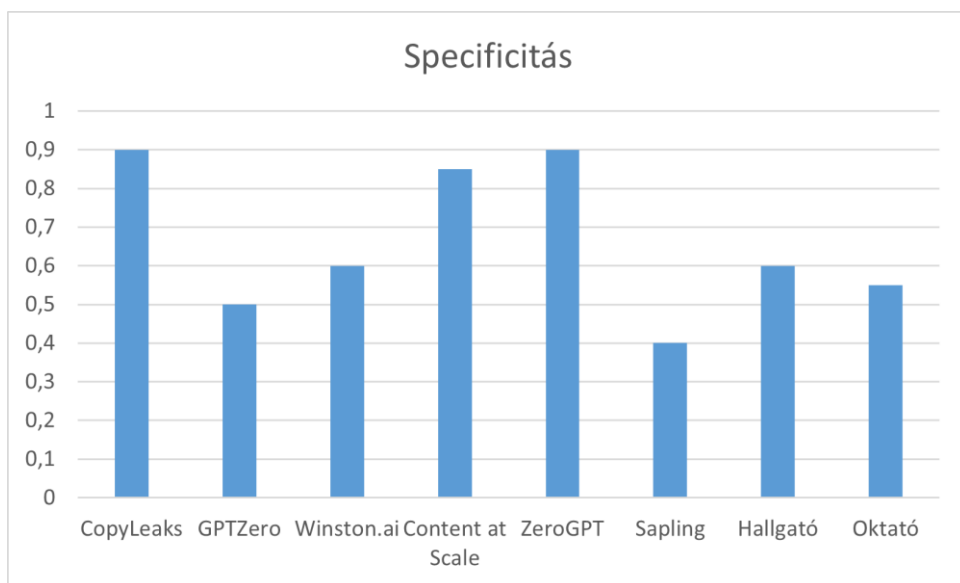
ábra 1: Az értékelőkhöz tartozó pontosság (saját szerkesztés)

A szenzitivitási értékeket bemutató 2. ábrán látható, hogy az értékelők jobban teljesítettek az AI által generált szövegek észlelésében. A szoftverek közül három (Winston.ai, Content at Scale, Sapling) minden AI által írt szöveget helyesen megtalált, a GPTZero 95, a CopyLeaks 80, a ZeroGPT pedig 70 százalékban adott helyes választ. Az emberi értékelők oldaláról az oktató 75, a hallgató pedig 55 százalékos eredményt ért el, így itt az oktató felülmúlta a ZeroGPT pontosságát.



ábra 2: Az értékelőkhöz tartozó szenzitivitás (saját szerkesztés)

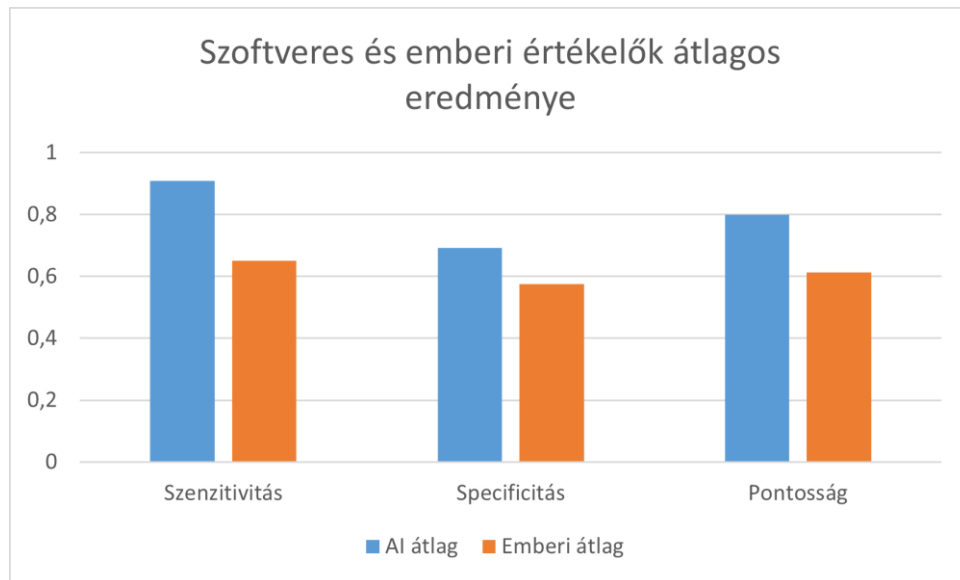
Az egyes értékelők specifikitása a 3. ábrán található. Itt rosszabb eredmények születtek, a 90 százalékot elérő CopyLeaks és ZeroGPT, illetve a 85 százalékos sikeraránnal válaszoló Content at Scale után a többi értékelő 60% alatt végzett. A Winston.ai és a hallgató 60, az oktató 55, a GPTZero 50, a Sapling pedig 40 százalékot ért el.



ábra 3: Az értékelőkhöz tartozó specifikitás (saját szerkesztés)

Elmondható, hogy az AI által generált szövegek megtalálása pontosabban ment az értékelőknek, átlagosan 18 százalékkal jobb eredmények születtek a szenzitivitás terén. Míg az AI által generált szövegek esetén 85%-ban helyes válasz született (160-ból 135 alkalommal), addig az emberek által írt szövegeket mindössze az esetek kétharmadában értékelték helyesen (160-ból 106 alkalommal).

Ahogy az a 4. ábrán is látható a szoftveres ellenőrzők szenzitivitása átlagosan 90, specificitása pedig 69 százalék. Az emberi értékelőknél ez 65 és 57,5 százalék. A szoftverek összesített pontosságának átlaga 80, az emberi értékelőké pedig 61,25 százalék. A különbség a pontosságnál 18,75, a szenzitivitásnál 35, a specificitásnál pedig 11,5 százalék.



ábra 4: Szoftveres és emberi értékelők átlagos eredményének összehasonlítása (saját szerkesztés)

Elmondható tehát, hogy az AI detektor szoftverek működnek, és ugyan az egyes típusok eredményeiben lényeges eltérések tapasztalhatóak, az emberi képességet meghaladó pontossággal tudják azonosítani egy szöveg származását. Azonban kiemelendő, hogy a legpontosabb eredményeket elérő szoftverek esetében is több tévesen pozitív válasz született a húsz darab emberek által írt szöveg értékelésekor. Ez az arány semmiképpen nem elhanyagolható, és hallgatók téves meggyanúsításához vezethet.

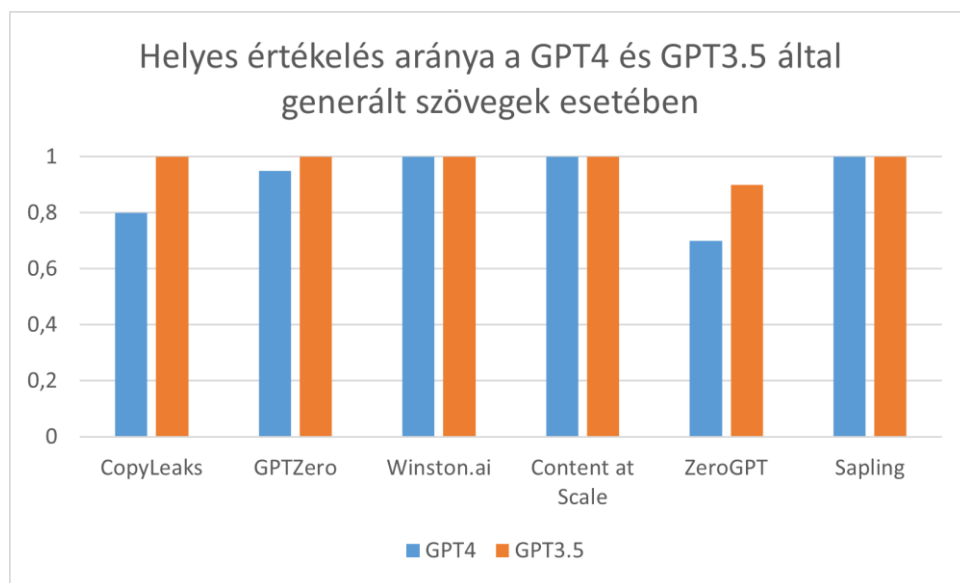
Természetesen, ahogy azt a módszertani részben is kiemelttem, a tesztelés rövid szövegrészekkel és viszonylag kis mintán zajlott. Nem zárható ki, hogy egy több, hosszabb szövegen alapuló tesztben jobban teljesítenének az ellenőrző szoftverek, akár elérve a készítőik által hirdetett pontosságot.

3.2. GPT3.5 ÉS GPT4 MODELLEK ÁLTAL GENERÁLT SZÖVEGEK ÖSSZEHASONLÍTÁSA

Az előző szakaszban is látszott, hogy az AI-generált szövegeket megbízhatóbban azonosítják az értékelő szoftverek, mint az emberek által írtakat. A GPT4 azonban egy viszonylag új és fejlett rendszer, ami előfizetéssel érhető el. Így vizsgálatra érdemes kérdés az is, hogy eredményesebben azonosítják-e az ingyenesen elérhető, korábbi modell, azaz a GPT3.5 által írt szövegeket.

A 4. ábrán láthatóak az egyes programok GPT3.5 és GPT4 által generált írásokra adott értékelései. A GPT4 esetében sem jellemző a sok hiba, három szoftver mind a húsz szöveget pontosan azonosította (Winston.ai, Content at Scale, Sapling), és a ZeroGPT követett csak el kiemelkedően sok hibát (hat szöveget nem azonosított AI-generáltként). A két másik program, a CopyLeaks és a GPTZero 4, illetve 1 AI-generált szöveget tekintett ember által írtnak.

A GPT3.5 esetében a hibák lényegében eltűnnek, egyedül a ZeroGPT hibázott két szöveg esetében. Ezeket azonban párban kapta, így itt is valójában egy esetben született téves eredmény.



ábra 5: Ellenőrző szoftverek pontossága a GPT4 és GPT3.5 modellek által generált szövegek értékelésekor (saját szerkesztés)

Összességében elmondható, hogy az AI detektorok a GPT4 által generált szövegekre is fel vannak készülve, azokat nagy pontossággal azonosítani tudják, de esetenként előfordulhat, hogy nem észlelik azokat. A GPT3.5 esetén erre még kisebb az esély, láthatóan segíti az észlelést, hogy a modell régebb óta elérhető és kevésbé komplex, kiszámíthatóbb.

4. ÖSSZEGZÉS

A mesterséges intelligencia 2022. végétől kezdődően a korábbiakhoz képest sokkal több embert ért el és kezdett el foglalkoztatni. Ez elsődlegesen az emberi nyelvhasználatot nagy pontossággal szimulálni tudó nagynyelvi modelleken keresztül jelent meg, melyek közül úttörő volt a korábbi rendszereknél nagyobb tudással bíró GPT3.5 modell, melynek elterjedését segítette az arra épülő ingyenesen használható és felhasználóbarát chatbot, a ChaGPT.

Az élet számtalan területén hasznosnak bizonyuló program az oktatási szférában is népszerűsége tett szert, felmérések szerint az egyetemi hallgatók nagyjából harmada használta már a programot valamilyen házi feladathoz vagy beadandó dolgozathoz.

Az AI eszközök ilyen jellegű használata csalásnak, akár plágiumnak is számíthat, és több módszerrel próbálnak védekezni ellen az oktatási intézmények. Egyes egyetemek elérhetlenné teszik a belső hálózatukon vagy kimondottan betiltják annak használatát a hallgatók körében. Azonban nehézséget jelent a használat észlelése, ugyanis a szoftverek által generált szövegeket ténylegesen nem könnyű megkülönböztetni az ember által írottaktól.

Ennek megoldására több AI detektor szoftver is megjelent, melyek önbevallásuk szerint jellemzően 95 százalék feletti pontossággal azonosítják, ha egy szöveg forrása valamilyen AI modell. Azonban az interneten rendre jelennek meg arról szóló posztok és hírek, hogy valakit tévesen gyanúsítottak meg AI eszközök igénybevételével és emiatt túlzó következményekkel néznek szembe.

Tanulmányomban az elérhető AI detektor szoftverek hatékonyságát vizsgáltam meg 20 darab AI-generált, és 20 darab emberek által írt szövegrészleten keresztül. Összehasonlításképpen emberi értékelőkkel, egy egyetemi oktatóval és egy hallgatóval is ellenőriztettem a szövegeket. Ezen túlmenően megvizsgáltam azt is, hogy könnyebben észlelhetőek-e GPT4-nél régebbi nyelvi modellel generált szövegek.

A kutatás rendelkezik korlátokkal, melyek negatívan befolyásolhatták az eredményt. A vizsgálat alapjául szolgáló minta mindössze negyven szöveg (illetve további húsz a GPT4 és GPT3.5 észlelhetőségének összehasonlításához), és azok csak egy bekezdés hosszúak. A detektor szoftverek pontossága saját bevallásuk szerint a vizsgált szövegek hosszának növelésével javul.

Az összes AI detektor szoftverről elmondható, hogy legalább a szövegek 70 százalékának forrását eltalálták. A legjobban a Content at Scale rendszere teljesített a vizsgálatban, ez 92,5 százalékos pontossággal rendelkezett. A szoftverek az AI-generált szövegeket majd minden esetben azonosítani tudták, a problémát az jelentette, hogy több esetben emberek által írt szövegeket is AI-generáltnak minősítettek.

Ez komoly problémát jelent, hiszen ezeket a detektor programokat sok esetben egyetemi oktatók használják a hallgatók munkájának ellenőrzésére. A téves pozitív jelzések hallgatók téves meggyanúsítását eredményezhetik, melynek következtében elbukhatják a tárgyat, vagy akár ennél komolyabb következményekkel is szembenézhetnek.

Az emberi értékelők mindketten rosszabbul teljesítettek, mint a legrosszabbul szereplő szoftver. Az oktató a szövegek 65 százalékát, a hallgató pedig 57,5 százalékát azonosította helyesen.

Bár a GPT4 által generált szövegeket is nagy pontossággal tudták észlelni a szoftverek, a GPT3.5 írásaival még eredményesebbek voltak, egy szoftver kivételével hiba nélkül sikerült azonosítaniuk azokat.

Egyesek szerint az AI-generált tartalmak detektálására tett kísérletek reménytelenek, hosszútávon a generatív AI modellek fejlődésével lehetetlen lesz azokat megkülönböztetni az emberek által készítettektől. Azonban a detektor szoftverek is folyamatosan fejlődnek, ahogyan az látszik is a korábbi kutatásokban. A nagy kérdést itt az jelenti tehát, hogy melyik oldal fejlődik gyorsabban, a detektorok, avagy a generatív modellek.

Teoretikusan van rá lehetőség a „watermarking”-nak nevezett módszerrel, hogy a nyelvi modellek működése egy olyan algoritmus szerint történjen, melynek ismeretében egyértelműen beazonosíthatóak az általa létrehozott szövegek. Ez amellett, hogy a biztos észlelést lehetővé teszi, a téves pozitív jelzéseket is kiküszöböli. Ezzel kapcsolatban azonban kérdés, hogy a modellek készítőinek érdeke-e az ilyen jellegű felismerhetőség.

Egy biztos, már rövidtávon is várhatóak fejlemények. Azt híresztelik, hogy már 2024. nyarán megjelenhet az OpenAI következő modellje, a GPT5. Elmondásuk szerint nagyjából akkora ugrást jelent majd a GPT4-hez képest, mint amekkorát az jelentett a GPT3 után. (Edwards, 2024)

IRODALOMJEGYZÉK

Szakirodalom

- Casal, J.E. & Kessler M. (2023). Can linguists distinguish between ChatGPT/AI and human writing?: A study of research ethics and academic publishing. *Research Methods in Applied Linguistics*. 2(3). <https://doi.org/10.1016/j.rmal.2023.100068>
- Crawford, J., Cowling, M., & Allen, K. (2023). Leadership is needed for ethical ChatGPT: Character, assessment, and learning using artificial intelligence (AI). *Journal of University Teaching & Learning Practice*, 20(3). <https://doi.org/10.53761/1.20.3.02>
- Elkhatat, A.M., Elsaid, K. & Almeer, S. (2023). Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text. *International Journal for Educational Integrity*, 19(17). <https://doi.org/10.1007/s40979-023-00140-5>
- Feuerriegel, S., Hartmann, J., Janiesch, C. & Zschech, P. (2023). Generative AI. *Business & Information Systems Engineering*. 66, 111-126. <https://doi.org/10.1007/s12599-023-00834-7>
- Fu, Y., Xiong, D. & Dong, Y. (2023). Watermarking Conditional Text Generation for AI Detection: Unveiling Challenges and a Semantic-Aware Watermark Remedy. *arXiv*. <https://doi.org/10.48550/arXiv.2307.13808>
- Gegg-Harrison, W. & Quarterman, C. (2024). AI Detection's High False Positive Rates and the Psychological and Material Impacts on Students. In Mahmud, S. (Ed.), *Academic Integrity in the Age of Artificial Intelligence*. pp. 199-219. IGI Global. <https://doi.org/10.4018/979-8-3693-0240-8.ch011>
- Krishna, K., Song, Y., Karpinska, M., Wieting, J. & Iyyer, M. (2023). Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense. *arXiv*. <https://doi.org/10.48550/arXiv.2303.13408>
- Pegorano, A., Kumari, K., Fereidooni, H. & Sadeghi, AR. (2023). To ChatGPT, or not to ChatGPT: That is the question! *arXiv*. <https://doi.org/10.48550/arXiv.2304.01487>
- Solaiman, I., Brundage, M., Clark J., Askell, A., Herbert-Voss, A., Wu, J., Radford, A., Krueger, G., Kim, J.W., Kreps, S., McCain, M., Newhouse, A., Blazakis, J., McGuffie, K. & Wang, J. (2019). Release Strategies and the Social Impacts of Language Models. *arXiv*. <https://doi.org/10.48550/arXiv.1908.09203>
- Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltynnek, T., Guerrero-Dib, J., Popoola, O., Sigut, P. & Waddington, L. (2023). Testing of detection tools for AI-generated text. *International Journal for Educational Integrity*, 19(26). <https://doi.org/10.1007/s40979-023-00146-z>
- Xu, Z. & Sheng, V.S. (2024). Detecting AI-Generated Code Assignments Using Perplexity of Large Language Models. *Proceedings of the AAAI Conference on Artificial Intelligence*. 38(1).

Források

- Chakravarti, A. (2023). Oh god! Open AI tool that identifies text written by ChatGPT believes Bible was written by AI. *India Today*. 2023. február 2. <https://www.indiatoday.in/technology/news/story/oh-god-open-ai-tool-that-identifies-text-written-chat-gpt-believes-bible-was-written-by-ai-2329163-2023-02-01> (Utolsó letöltés: 2024.04.09.)

- Content at Scale (2022). *AI Content Detector*. <https://contentatscale.ai/ai-content-detector/> (Utolsó letöltés: 2024.04.10.)
- CopyLeaks (2023a). *AI Content Detector*. <https://copyleaks.com/ai-content-detector> (Utolsó letöltés: 2024.04.10.)
- CopyLeaks (2023b). *AI Content Detector Languages*. <https://copyleaks.com/wp-content/uploads/2023/09/Copyleaks-ai-content-detector-languages.pdf> (Utolsó letöltés: 2024.04.10.)
- Edwards, B. (2024). GPT-5 might arrive this summer as a “materially better” update to ChatGPT. *Ars Technica*, 2024. március 20. <https://arstechnica.com/information-technology/2024/03/openais-gpt-5-may-launch-this-summer-upgrading-chatgpt-along-the-way/> (Utolsó letöltés: 2024.04.07.)
- Foote, K.D. (2023). A Brief History of Large Language Models. Dataversity, 2023. december 28. <https://www.dataversity.net/a-brief-history-of-large-language-models/> (Utolsó letöltés: 2024.04.14.)
- GPTZero (2023). Homepage. <https://gptzero.me> (Utolsó letöltés: 2024.04.10.)
- Henry, M. (2023). Nearly a third of college students used ChatGPT last year, according to survey. *Ohio Capital Journal*. 2023. szeptember 25. <https://ohiocapitaljournal.com/2023/09/25/nearly-a-third-of-college-students-used-chatgpt-last-year-according-to-survey/> (Utolsó letöltés: 2024.04.09.)
- Hu, K. (2023). ChatGPT sets record on fastest growing user base - analyst note. *Reuters*, 2023. február 2. <https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/> (Utolsó letöltés: 2024.03.19.)
- Kirchner, J. H. (2023). New AI classifier for indicating AI-written text. *OpenAI Blog*, 2023. január 31. <https://openai.com/blog/new-ai-classifier-for-indicating-ai-written-text> (Utolsó letöltés: 2024.04.07.)
- Mitchell, A. (2023). AI thinks the Constitution was written by bots — but there’s a bigger issue. *New York Post*, 2023. július 25. <https://nypost.com/2023/07/25/why-its-a-problem-that-ai-thinks-the-constitution-was-made-by-ai/> (Utolsó letöltés: 2024. 04.09.)
- O’Neill, S. (2023). The History of OpenAI. *Lxahub.com*, 2023.05.02. <https://www.lxahub.com/stories/the-history-of-openai> (Utolsó letöltés: 2024. 04.15.)
- Sapling (2023). *AI Content Detector*. <https://sapling.ai/ai-content-detector> (Utolsó letöltés: 2024.04.10.)
- Winston.ai (2023). Homepage. <https://gowinston.ai/> (Utolsó letöltés: 2024.04.10.)
- ZeroGPT (2023). Homepage. <https://www.zerogpt.com/> (Utolsó letöltés: 2024.04.10.)

