



M Ű E G Y E T E M 1 7 8 2

BUDAPESTI MŰSZAKI ÉS GAZDASÁGTUDOMÁNYI EGYETEM

TDK DOLGOZAT

Mesterséges intelligencia a matematikában: A ChatGPT képességei és gyakorlati hatásai

MÁTÉ BENDE

ALTERNATÍV KÖZGAZDASÁGI GIMNÁZIUM 12. ÉVFOLYAM

TÉMAVEZETŐ:

Molontay Roland

EGYETEMI DOCENS, BME, SZTOCHASZTIKA TANSZÉK

Pintér József

PHD HALLGATÓ, BME, SZTOCHASZTIKA TANSZÉK

2024

Tartalomjegyzék

1. Bevezetés	2
2. Irodalomáttekintés	3
2.1. Nagy nyelvi modellek és matematika a tudományos irodalomban	3
2.2. Interjúk Terence Taoval	5
3. ChatGPT matematikában	7
3.1. Nagy nyelvi modellek	7
3.2. ChatGPT használata	10
3.3. ChatGPT magyar és angol nyelven	15
3.4. Matematika Érettségi	15
3.5. Eredmény témakörönként	20
3.6. Tipikus hibák	21
3.7. KöMaL	27
4. Kérdőív	32
4.1. A kitöltők szokásai, tapasztalatai és véleménye a ChatGPT-vel kapcsolatban . .	32
4.2. Definíciók vizsgálata - ChatGPT mint tanár	36
4.3. Kitekintés - ChatGPT a jövőben	40
5. Zárógondolatok	43

1. Bevezetés

A mesterséges intelligencia (MI) technológiájának dinamikus és gyors ütemű fejlődése az elmúlt évtizedek során számos iparágban és különböző területeken alapvető átalakulásokat idézett elő, amelyek közül az oktatás is kiemelt figyelmet kapott. Különösen a nagy nyelvi modellek, például a ChatGPT jelentős előrelépést hoztak a tanulási és tanítási folyamatokban, új perspektívákat nyitva meg mind a diákok, mind az oktatók számára. A ChatGPT működésének alapját a hatalmas mennyiségű szöveges adat feldolgozása és a gépi tanulási algoritmusok alkalmazása képezi, amely lehetővé teszi számára, hogy természetes nyelven folytasson kommunikációt, valamint komplex matematikai problémákat értelmezzen és hatékonyan megoldjon. Ezen képességek révén a ChatGPT nem csupán egyszerű megoldásokat kínál a felmerülő kérdésekre, hanem részletes és érthető magyarázatokat is nyújt, ezzel jelentősen támogatva a diákok tanulási folyamatait és elősegítve a mélyebb megértést.

Az oktatási szektorban számos kihívás jelentkezik nap mint nap előtérbe, például a személyre szabott oktatás iránti növekvő igény és a tanárok egyre nagyobb terheltsége, amelyek megoldása érdekében innovatív és hatékony megoldásokra van szükség. A ChatGPT és hasonló MI-alapú eszközök lehetőséget kínálnak ezeknek a problémáknak a kezelésére azáltal, hogy személyre szabott segítséget nyújtanak a diákoknak, valamint azonnali visszajelzéseket biztosítanak a tanulási folyamat során. Ezáltal a tanulók nemcsak egyéni szükségleteihez igazodó támogatást kapnak, hanem a tanárok munkaterhelése is csökkenhet, mivel az MI képes automatizálni bizonyos feladatokat és hatékonyabbá tenni a tanítási módszereket.

Kutatásom elsődleges célja annak részletes vizsgálata, hogy a ChatGPT milyen mértékben képes teljesíteni különböző matematikai feladatsorokat, beleértve a közép- és emelt szintű érettségi vizsgák feladatait, valamint a KöMaL (Középiskolai Matematikai és Fizikai Lapok) matematika versenyének A, B és C kategóriás feladatait. A kutatás során azt elemzem, hogy a modell hogyan viselkedik a különböző matematikai témakörökben, mint például az algebra, geometria vagy a kombinatorika, és hogy ezekben a területeken milyen mértékben képes helyes és pontos megoldásokat nyújtani. Emellett azt is vizsgálom, hogy a feladatok nyelve – legyen az magyar vagy angol – hogyan befolyásolja a ChatGPT teljesítményét, ezzel is hozzájárulva annak megértéséhez, hogy a nyelvi kontextus milyen szerepet játszik az MI hatékonyságában.

A kétrészes kutatásom második eleme egy kérdőív, mely segítségével fel tudom térképezni a diákok szokásait és véleményét a ChatGPT-ről matematikával kapcsolatban. Tovább vizsgálom azt is, hogy a diákok mennyire részesítik előnyben a ChatGPT által generált definíciókat, valamint, hogy milyen mértékben ismerik fel a "gépi" eredetű tartalmakat.

A kutatásom fő hipotézise az, hogy a ChatGPT képes helyesen megoldani magas szintű matematikai feladatokat, és hogy teljesítménye nyelvi szempontból függetlenül hasonló eredményeket ér el. Úgy vélem, hogy a ChatGPT teljesítménye jelentősen eltér a különböző matematikai témakörökben, például az algebra és a geometria területén, ami fontos tényező lehet az oktatási alkalmazások fejlesztése során. Továbbá azt feltételezem, hogy a diákok a ChatGPT által

generált „egyszerűbb” definíciókat találják a legérthetőbbnek, ugyanakkor a többségük nem ismeri fel, hogy ezeket a definíciókat a modell hozta létre, ami rávilágíthat arra, hogy az MI által nyújtott tartalmak mennyire tudnának integrálódni a tanulási folyamatba nagyobb zavar nélkül.

A kutatás során a ChatGPT által megoldott feladatokat kézi ellenőrzéssel értékeltem, összehasonlítva azokat a hivatalos megoldókulcsokkal. Ez a módszer biztosítja, hogy a modell által nyújtott megoldások pontosságát és megbízhatóságát alaposan meg lehessen vizsgálni. Kutatásom egyedülálló abban, hogy részletesen elemzi a modell teljesítményét különböző feladattípusokban és nyelvi kontextusokban, de a legfontosabb, hogy integrálja a diákok perspektíváját, ami átfogó képet nyújt az MI oktatási potenciáljáról.

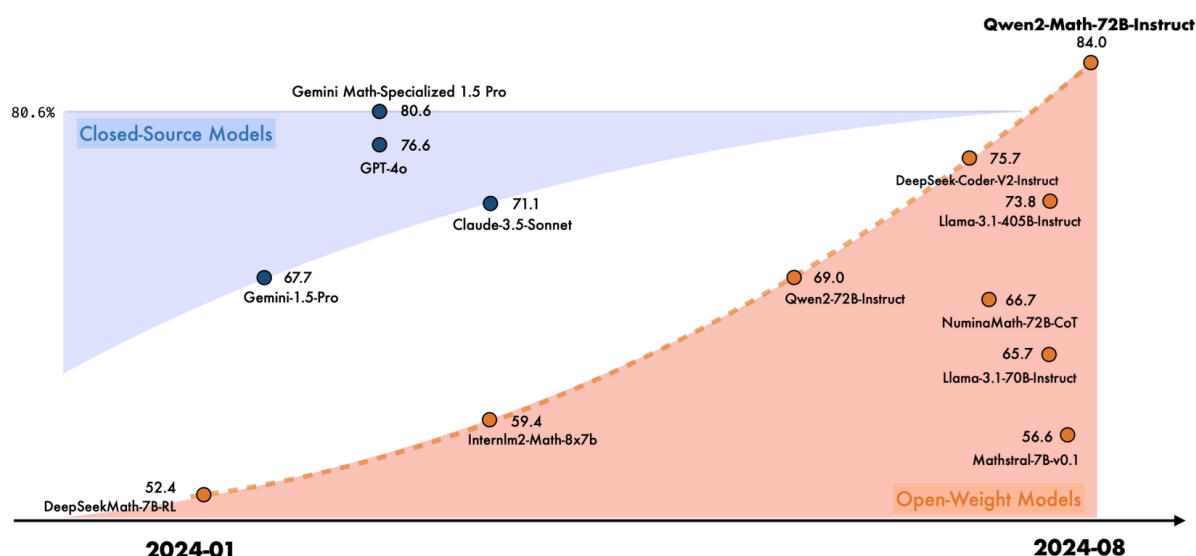
Összességében úgy vélem, hogy a ChatGPT jelentős potenciállal rendelkezik a matematika oktatásában, és kutatásom fontos lépést jelenthet a technológia hatékony integrálása felé az oktatási folyamatokba. A mesterséges intelligencia alkalmazása az oktatásban nemcsak a tanulási módszerek modernizálását teszi lehetővé, hanem hozzájárulhat ahhoz is, hogy a diákok mélyebb és átfogóbb tudásra tegyenek szert, miközben az oktatók munkáját is támogatja és könnyíti. Ezzel a kutatással reményeim szerint hozzájárulok ahhoz, hogy az MI technológiák optimálisan kerüljenek beépítésre az oktatási rendszerekbe, elősegítve ezzel a jövő generációinak sikeres és hatékony tanulását.

2. Irodalomáttekintés

Ebben a fejezetben ismertetek néhány alapvető tudományos munkát nagy nyelvi modellek és a matematika kapcsolatáról, illetve bemutatok két tanulságos interjút korunk egyik legnagyobb, ha nem a legnagyobb matematikusával, Terence Taoval.

2.1. Nagy nyelvi modellek és matematika a tudományos irodalomban

Semmi sem mutatja jobban a nagy nyelvi modellek és a matematika kapcsolatára vonatkozó kutatások fontosságát, minthogy a világ legnagyobb informatikai vállalatai dollármilliárdokat fektetnek különböző, akár matematikára specializálódott nagy nyelvi modellek fejlesztésébe. Az egyik legnagyobb zárt forráskódú modell a Gemini Math-Specialized 1.5 Pro (Gemini Team Google (2024)), melyet a Google fejlesztett, és egészen 2024 augusztusáig, a Qwen2-Math-72B-Instruct modell megjelenéséig, a lehető legerősebb matematikára specializálódott nagy nyelvi modellként tartották számon (Qwen Team (2024)) (lásd 1. ábra). Míg a Gemini egy zártforráskódú modell, addig az őt 2024 augusztusában leköröző Qwen2 modell már nyíltforráskódú, amely azt jelenti, hogy bárki szabadon hozzáférhet a teljes modellhez, melynek minden egyes apró részlete nyilvános. Ennek következtében a Qwen2 egy különösen figyelemreméltó modell, hiszen nemcsak kiválóan teljesít az eddigi state-of-the-art modellhez képest, hanem utat is nyit rengeteg továbbfejlesztési lehetőségre. Azonban mitől lesz egy modell jobb a többinél? Természetesen a legfőbb befolyásoló tényező a mögöttes adathalmaz, azonban léteznek



1. ábra. Matematikára specializálódott és általános nagy nyelvi modellek teljesítménye a GSM8K adathalmazon, mely angol nyelvű szöveges jellegű matematikai feladatokat tartalmaz változó nehézségben. Az ábrán megfigyelhető, hogy az új áttörések a matematika területén nyílt forráskódú modellekkel születnek, melyek már a Google Gemini és az OpenAI GPT-4o modelljét is lekörözik. *Kép forrása: Qwen2 – <https://qwen2.org/math/>*

különböző technikák, melyekkel jelentősen növelni lehet a nagy nyelvi modellek teljesítményén különösen olyan feladatköröknél, ahol elengedhetetlen a precíz érvelés.

Az egyik ilyen fontos terület a felhasználók és a modellek (vagy a modellek és a modellek) közötti interakció, amelyet gyakran „prompting”-nak neveznek. A SciEx (Dinh, Mulloy, Bärman, and et al. (2024)) értékelő rendszer azt vizsgálja, hogy az LLM-ek milyen jól teljesítenek tudományos vizsgákon, beleértve olyan nyílt végű kérdéseket is, amelyek még mélyebb érvelést igényelnek. A kutatás azt találta, hogy mindenféle egyéb trükkök nélkül még a legfejlettebb modellek is küzdenek a komplex, bonyolult matematikai érvelésekkel, azonban promptolási technikák segítségével jelentősen növelhető a modellek teljesítménye. A legegyszerűbb promptolási technikára példa a chain-of-thought promptolás, mely a modellt arra biztatja, hogy a feladatokat részfeladatokra bontva, lépésről lépésre oldják meg. Imani és Shrivastava (Imani, Du, and Shrivastava (2023)) cikke szerint ezen egyszerű promptolási technika akár 60%-kal is növelheti a modellek matematikai teljesítményét aritmetikai jellegű problémák megoldásában (MultiArith adathalmaz).

Kérdéses, hogy milyen magas szintre lehet eljutni, ha rendelkezésre állnak mind megfelelő adatok, és promptolási technikákat is a legmagasabb szintre emeljük? Erre példa a Google AlphaGeometry (Trinh, Wu, Le, He, and Luong (2024)) modellje, amely képes olyan magas szintű geometriaproblémákat megoldani, mint amilyenek a matematikai diákolimpiákon találhatóak. Ráadásul az AlphaGeometry nem támaszkodik meglévő emberi megoldásokra; helyette neurális hálózattal és szimbolikus érveléssel (mint például a Coq vagy Lean matematikai bizonyítók, bizonyítás ellenőrző rendszerekben találhatóak) generál szintetikus adatokat a tanulásához.

Ezzel az eljárással az AlphaGeometry oly magas szintre emelte a nagy nyelvi modellek matematikai teljesítményét, hogy a modelljük jobban teljesít egy átlagos ezüstérmes versenyzőtől a Nemzetközi Matematika Olimpián (IMO).

De hova vezet az, ha a nagy nyelvi modellek messze túlszárnyalják az átlagos ember kognitív teljesítményét nemcsak a matematikában, hanem hétköznapi teendőkben is? Mivel az LLM-ek egyre erősebbé és szélesebb körben használttá válnak, és így egyre inkább támaszkodunk rájuk, aggodalmak merülnek fel az emberi kognitív képességekre gyakorolt hatásuk kapcsán. Richard Heersmink (Heersmink (2024)) egyik cikke szerint az LLM-ek hatással lehetnek kognitív képességeinkre azáltal, hogy átveszik az olyan gondolkodást igénylő feladatokat, amelyek kritikai gondolkodást igényelnek, mint például a matematikai problémamegoldás, bizonyítás, vagy akár egy egyszerű e-mail megírását. Heersmink szerint, ha túlzottan támaszkodunk az LLM-ekre, bizonyos készségeink elveszhetnek, hasonlóan ahhoz, ahogyan a GPS használata csökkentheti a térbeli tájékozódásunkat.

A mesterséges intelligencia oktatásban betöltött szerepéről sem feledkezhetünk meg. C. K. Lo cikkében (Lo (2023)) 50 tudományos munkát elemzett ki AI és oktatás kapcsolatáról. Munkájában beszámol arról, hogy milyen különböző alkalmazási módszerekkel használnak a világ különböző országaiban AI-t oktatásra, és hogy milyen problémák merülnek fel, és azokra milyen megoldások adhatóak. Az egyik legalapvetőbb kritikus kérdés, hogy egyáltalán képes-e az AI megbízható, helyes válaszokat generálni a kérdésekre. Ezen kérdés megválaszolásával foglalkozok én is ezen dolgozatomban, melyben bemutatom, hogy a jelenlegi legismertebb nyilvánosan elérhető mesterséges intelligencia, a ChatGPT kiválóan érti a közép- és emelt szintű középiskolás matematikát, sőt egy bizonyos szintig még a komolyabb versenyfeladatokat is. Azonban abban az esetben, ha valamilyen adott területen az AI mégsem teljesítene jól, lehetséges azokat finomhangolni egy-egy specifikus tudományterületre ("Google Cloud: Dialogflow" (2023); Topsakal and Topsakal (2022)). Amennyiben viszont az AI kiválóan értene egy-egy tudományterülethez, például matematikához, akkor az rengetegféleképpen hasznosítható az oktatásban, például segítheti a hallgatókat a tanulásban a tananyag elmagyarázásával, feladatok megoldásának ellenőrzésével, a hibákra való rávilágításra és azok megmagyarázására, a feladatok helyes megoldásával, mindezt úgy, hogy bármikor rendelkezésre állva. Sőt a tanárok munkáját is megkönnyíthetné, például feladatok kitalálásával, vagy az adott tematikához tartozó teljeskörű tananyag, jegyzet generálásával (Lo (2023)).

2.2. Interjúk Terence Taoval

Míg a tudományos irodalom adatokkal és tényekkel támasztja alá a nagy nyelvi modellek matematikai teljesítményét, addig ezen fejezetben Terence Tao szubjektív, de jól megalapozott véleményét mutatjuk be.

Terence Tao Fields-érem nyertes matematikus több interjút is adott az elmúlt hónapokban nagy nyelvi modellek és matematika kapcsolatáról. Scientific Americannak adott interjújában (Tao (2024a)) kifejtette véleményét a témáról, melyet röviden összefoglalni úgy lehetne, hogy

az AI lesz a matematikus Co-Pilotja; de vajon mit is jelent ez? Ahhoz, hogy megértsük Terence Tao véleményét, meg kell ismernünk a bizonyítás segítő rendszereket. Manapság több bizonyítás segítő rendszer is létezik, melyek közül a legismertebbek a Lean, Coq vagy Isabelle. Ezen rendszerek működésének lényege, hogy a felhasználók, matematikusok különböző axiómákat fektethetnek le, majd azokból logikai összefüggések alapján állításokat, tételeket vezethetnek le. A bizonyítás segítő rendszer lényege, hogy automatikusan ellenőrzi, hogy a bizonyítás helyes-e. Ezen rendszerek manapság kevésbé elterjedtek, melynek alapvetően két oka van, az egyik, hogy a közhiedelemben elterjedt, hogy ezen rendszerek csak nagyon alapvető egyszerű állítások bizonyítására képesek, illetve a másik, hogy használatuk kifejezetten nehezen elsajátítható, mély tudást igényel, mint szintaktikai, mind szemantikai szinten. Terence Tao ezen interjújában erre a két problémára is reflektál, megemlíti egy Leanben programozott csomagot, mely már az egész BSc szintű matematikai tudást tartalmazza tétel, bizonyítás szinten. Továbbá a nehezen elsajátíthatóságnál jönnek képbe az AI rendszerek, melyekről ma már köztudott, hogy kiválóan tudnak programozni. Ezen rendszerek nemcsak Python nyelven, de akár Lean nyelven is képesek egyre jobban programozni, ezáltal segíthetik a matematikust többek között a bizonyításaik ellenőrzésében. Ebben látja Terence Tao az AI és matematika jövőjét, az AI egy olyan segédpilóta (Co-Pilot) lehet, mely bizonyítás ellenőrző rendszerekben különböző tételek, állítások leprogramozásában nyújt segítséget ezáltal megbizonyosodva a bizonyítások helyességéről. Tao szerint ez a matematikusok közötti együttműködést is jelentősen meg tudja könnyíteni, hiszen, ha a bizonyítás valamilyen bizonyítás ellenőrző nyelven is meg van írva, akkor nem szükséges külön ellenőrizni egymás munkáját, hanem biztosak lehetünk benne, hogy kollégáink munkája helyes. Megjegyezzük, hogy az AlphaGeometry szintén Lean programnyelv és AI összekapcsolásával épített olyan nagy nyelvi modellt, mely ezüstérmet ért a nemzetközi matematika olimpián, ezzel többek között megerősítve Terence Tao sejtését is.

Tao egy másik, a The Atlanticnak adott interjújában (Tao (2024b)) a ChatGPT legújabb modelljét a GPT-o1 modellt hasonlítja össze matematikus hallgatókkal. A modell fejlesztői, az OpenAI azt állítják, hogy az új GPT-o1 modelljük PhD szintű tudással rendelkezik különböző területeken. Mivel jelenleg az o1 modell még fejlesztés alatt áll, így a pontos teljesítményéről egyelőre még nem állnak rendelkezésre számszerű adatok, azonban tapasztalatok alapján, Tao véleménye szerint a GPT-o1 egy átlagos mesterszakos hallgató szintjén áll jelenleg. Mit jelent ez a hallgatókra nézve? Mi a helyzet akkor, ha túlszárnyalja a hallgatók teljesítményét a ChatGPT, célszerű lehetne-e akkor hallgatók helyett AI-t alkalmazni a különböző kutatásokhoz? Tao szerint továbbra is célszerű AI helyett hallgatókat alkalmazni a különböző kutatásokhoz, melynek oka, hogy az AI modellek alapvetően statikusak, melynek jelentése, hogy ugyan valamennyire tanulnak a beszélgetésekből, de csak nagyon minimálisan. Ezzel ellentétben a hallgatóknak meg lehet tanítani új, akár mélyebb dolgokat is, melyek túlszárnyalják az adott területen az AI működését. Emellett Tao az interjújában azt is kiemeli, hogy jelentős javulást tapasztalt az GPT-o1 teljesítményében a GPT-4o-hoz képest.

3. ChatGPT matematikában

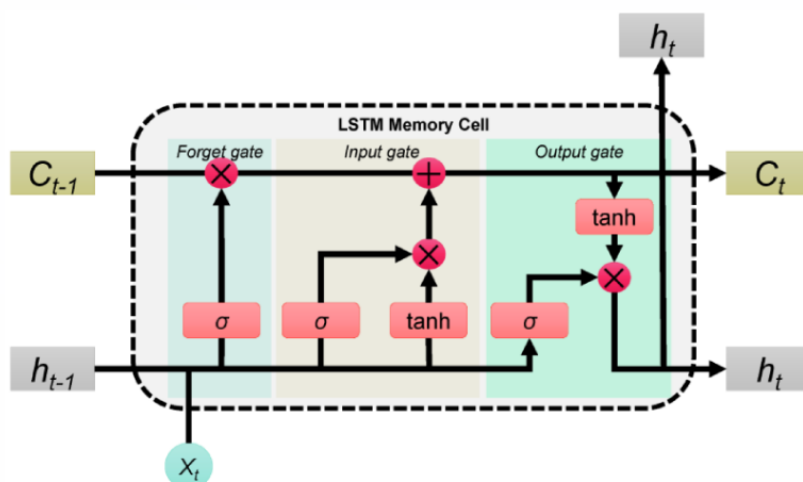
Mielőtt belekezdenék a ChatGPT matematikai képességének vizsgálatába, előtte bemutatom a legfontosabb tudnivalókat a nagy nyelvi modellekről és a ChatGPT használatáról.

3.1. Nagy nyelvi modellek

A mesterséges intelligencia (artificial intelligence, AI) és a gépi tanulás (machine learning, ML) fejlődése az elmúlt évtizedekben jelentős előrelépéseket hozott a természetes nyelv feldolgozásában (natural language processing, NLP). A nagy nyelvi modellek (Large Language Models, LLM-ek) olyan fejlett AI-rendszerek, amelyek képesek emberi nyelv megértésére és szövegek generálására. Ezek a modellek hatalmas mennyiségű szöveges adat feldolgozásával tanulnak, és számos alkalmazási területen, például fordításban, szövegösszefoglalásban és chatbotokban nyújtanak kiemelkedő teljesítményt.

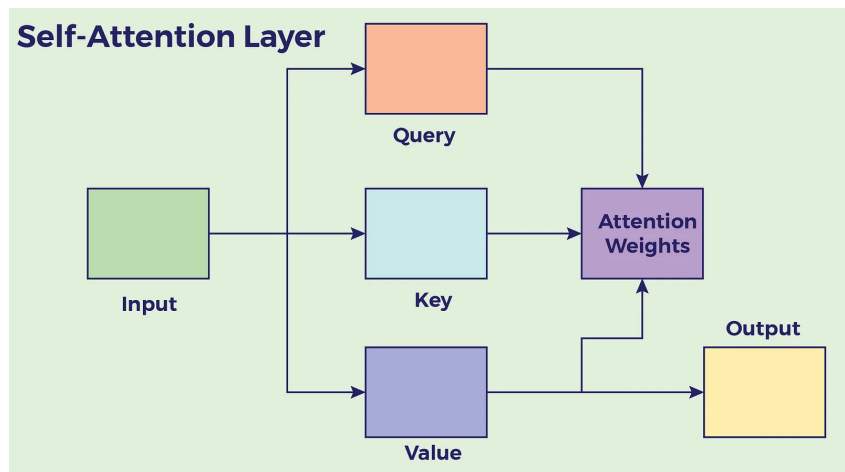
A nyelvi modellek fejlődése kezdetben statisztikai módszerekkel indult, ahol a szavak előfordulási gyakoriságát és együttjárását vizsgálták különböző korpuszokon. A 2010-es évek elején a neurális hálózatok alkalmazása forradalmasította a területet, különösen a visszacsatolt neurális hálózatok (recurrent neural networks, RNN) és azok továbbfejlesztett változatai, mint például a hosszú rövid távú memóriák (long short-term memory, LSTM, lásd x. ábra). Ezek a modellek, az egyszerű neurális hálózatokkal ellentétben, képesek voltak figyelembe venni a szövegben lévő szavak sorrendjét és kontextusát, ami javította a nyelvi feldolgozás pontosságát (Hochreiter and Schmidhuber (1997)).

Ugyan az LSTM alapú hálózatok kiváló teljesítményt nyújtanak, gyakorlatban nagyméretű adathalmazok feldolgozására alkalmatlanok, hiszen nem párhuzamosíthatóak, a szövegeket szóról szóra dolgozzák fel. Ennek a problémának a megoldására, a következő jelentős előrelépést a Transformer architektúra bevezetése jelentette 2017-ben, amely az "Attention Is All You Need" című tanulmányban került bemutatásra (Vaswani, Shazeer, Parmar, and et al. (2017)). A Transformer modellek az úgynevezett figyelem (attention) mechanizmust alkalmazzák, amely lehetővé teszi, hogy a modell a szöveg különböző részeire fókuszáljon, figyelembe véve a szavak közötti kapcsolódásokat és jelentésbeli összefüggéseket. Ez a megközelítés lehetővé tette a modelleknek a párhuzamos adatfeldolgozást, ami jelentősen felgyorsította a tanulási folyamatot, és hatalmas méretű adathalmazok feldolgozására adott lehetőséget.



2. ábra. A hosszú rövid távú memóriával rendelkező AI rendszerek, olyan rekurrens neurális hálózatok, melyeket természetes nyelvfeldolgozásban alkalmaznak. Működésükhöz három kapura van szükségük, melyeket felejtő (forget), bemeneti (input) és kimeneti (output) kapuknak hívnak. A neurális hálózat szóról szóra (x_t) beolvassa a feldolgozandó szöveget, majd minden egyes szónál a kapuk segítségével frissíti a memóriát (c_t) és az adott szóhoz tartozó állapotvektort (h_t). A felejtő kapu Hadamard-szorozattal kontrollálja, hogy a memória mely részét tartsa meg a neurális hálózat, a bemeneti kapu fogadja az új információkat a memóriába, majd a kimeneti kapu a memória állapota és az adott szó alapján meghatározza a szóhoz tartozó állapotvektort. *Kép forrása: ML Archive – Understanding Long Short-Term Memory (LSTM) Networks*

A Transformer modellek alapja a figyelem mechanizmus, amely lehetővé teszi, hogy a modell meghatározza, mely szavak fontosak egy adott szó értelmezéséhez. Ez különösen hasznos hosszú szövegek esetén, ahol a szavak közötti távoli kapcsolatok is relevánsak lehetnek. A figyelem mechanizmus három fő komponenst használ: lekérdezéseket (queries), kulcsokat (keys) és értékeket (values). Ezek segítségével a modell súlyozza a szavak közötti kapcsolatokat, és ennek alapján generálja a megfelelő kimenetet. A transformer architektúrák egyik további kulcsfontosságú eleme a multi-head attention réteg, amely lehetővé teszi, hogy a modell párhuzamosan több különböző nézőpontból is vizsgálja a szöveget. Ez növeli a modell képességét a komplex nyelvi mintázatok felismerésére és megértésére.



3. ábra. A transformer alapú neurális hálózatok magját képező figyelem (attention) mechanizmus. Az eljárás lényege, hogy a neurális hálózat nem szóról szóra dolgozza fel a bejövő mondatot, hanem egy adott vektortérbe leképezett, a mondat szavaihoz tartozó vektorok együttes mátrixát, egyszerre dolgozza fel mátrixszorzásokkal. A query mátrix a szavak kontextusát reprezentálja, míg a key mátrix azt, hogy mennyire relevánsak a különböző lekérdezésekre az egyes szavak. A query és key mátrixok szorzata megadja, hogy hol találjuk a releváns választ a mondatban a kérdésre, majd a value mátrixszal összehasonlítva megkapjuk a tényleges választ, információt. A működést legkönnyebben úgy érthetjük meg, ha egy egyszerűbb szituációba helyezzük a problémát, például keresőmotorok működésébe. Egy keresőmotor a mi lekérdezésünket (query), összehasonlítja az adatbázisban található értékekkel (key), majd az alapján, hogy melyik kulcs tartozik a lekérdezésünkhöz, a hozzátartozó találatot (value) adják vissza. *Kép forrása: ProjectPro – Demystifying Transformers Architecture in Machine Learning*

A ChatGPT szoftver alapja a GPT – Generative Pre-trained Transformer (OpenAI (2024)) neurális hálózat, mely egy olyan transformer alapú mesterséges intelligencia, mely generative: szövegek generálására lett fejlesztve és pre-trained: hatalmas korpuszon előre be lett tanítva, mielőtt különböző feladatokra, például chatbot szerep betöltésére finomhangolták volna.

3.2. ChatGPT használata

A ChatGPT az OpenAI nagy nyelvi modellje, mely az interneten elérhető chatbot és API formájában is. A chatbot a <https://chat.openai.com/> címen regisztrációt követően érhető el, és szövegesen közben lehet vele kommunikálni (lásd 4. ábra), míg az API használatához külön API-kulcsot kell igényelni és a hozzá tartozó modulok segítségével hívhatóak meg különböző programozási nyelveken, például Python nyelven.



4. ábra. A ChatGPT üdvözlőfelülete. A mezőbe különböző üzeneteket lehet írni, melyek lehetnek például matematikai feladatok, melyeket meg szeretnénk oldani, de a megfelelő gombokra kattintva akár képként is feltölthetőek, vagy a feladatot tartalmazó weboldal linkjét is megadhatjuk.

Jelenleg két modell áll rendelkezésre a felhasználók számára, a GPT-4o modell és a jelenleg is fejlesztés alatt álló GPT-o1 modell, melyek közül előbbi használata ingyenes, utóbbi használata pedig előfizetéshez kötött (bár az egyszerűbb o1-mini modellt előfizetés nélkül is ki lehet próbálni).¹ Azt hogy pontosan mi a különbség a különböző modellek között részben szakmai titok fedi, hiszen zártforráskódú modellekről van szó.² Azonban azt a tendenciát láthatjuk, hogy a modellek egyre nagyobbak és nagyobbak lesznek, akár 2 ezer milliárd paraméterrel is rendelkezhetnek, és egy komplexebb promptolási technikákat alkalmaznak a különböző programok belső kommunikációja során.

¹Megjegyzendő, hogy a ChatGPT folyamatosan fejlődik, a korábbi GPT-3.5 modellt felváltotta a GPT-4, majd pedig a GPT-4o modell, mely során azt tapasztalhattuk, hogy mindig egyre erősebb, és erősebb modellek kerülnek nyilvánosságra, és mindig az erősebb fizetős modell vált a későbbiekben ingyenessé, emiatt csak idő kérdése, hogy a GPT-o1 modell is mindenki számára hozzáférhető legyen.

²Az irodalomkutatásnál említettem, hogy léteznek nyílt forráskódú modellek, melyek mind a Google, mind az OpenAI régebbi modelljeinél jobban teljesítenek matematikai problémákon (az o1 pontos teljesítménye még nem ismert, de várhatóan túlszárnyalja az eddigi modelleket). Ezen modellek, mint például a Qwen2 ingyenesen elérhetőek bárki számára, azonban nagyon erős gép szükséges a futtatásukhoz, mely miatt ezeket nem tudtam kipróbálni, így kutatásom a két legismertebb, nyilvánosan elérhető modellre korlátozódik.

Különböző promptolási technikák segítségével jelentősen növelhetjük a modellek teljesítményét. Az egyik ilyen promptolási technika a chain-of-thought néven ismert, melynek lényege, hogy a felhasználó külön megkéri a chatbotot, hogy a feladatot részfeladatokra bontva, lépésről lépésre oldja meg. Egy másik ismert promptolási technika a few-shot prompting, melynek lényege, hogy a felhasználó megadja, hogy pontosan hogyan is szeretné várni a feladat megoldását, mely lehet például néhány hasonló feladat megoldása is. Általában azt a promptolási technikát, melyben nem alkalmazunk semmilyen útmutatást, zero-shot promptingnak szokás nevezni.

Vizsgáljunk meg egy egyszerű matematikai probléma megoldását ChatGPT-vel, és chain-of-thought promptolási technikával (lásd 1. példa). Megjegyzendő, hogy ezt a feladatot könnyedén megoldja a ChatGPT bármelyik verziója, azonban bonyolultabb feladatoknál már szükséges lehet komplexebb promptolási technikát is megadni. A ChatGPT-4o modellje képi formátumot és internetes linkeket is támogat input formájában, így könnyedén megoldatható akár egy egyetemi zh feladat is egyszerűen a feladatot screenshotolva (lásd 2. példa). Végül tekintsünk meg egy példát egy KöMaL feladatra is (lásd 3. példa). Versenyfeladatok esetében nemcsak az a fontos kérdés, hogy meg tudja-e korrektül oldani a feladatokat a nagy nyelvi modell, hanem az is, hogy mekkora mértékű segítséget tud nyújtani a versenyzők számára. Ebben a példában, ha a megoldás nem is tökéletes, bőven tartalmaz annyi hasznos észrevételt, amely jelentősen segítik a versenyzőt a feladat megoldásában, hiszen szinte tálcán kínálja a megoldás fő ötletét a ChatGPT.

Hol van a határ? Hol tart jelenleg a ChatGPT matematikai képessége? Erre a kérdésre keresem a választ a következő fejezetekben a tematikát középiskolai szintű matematikára szűkítve.

Segíts megoldani a következő feladatot. A feladatot lépésről lépésre oldd meg, és magyarázd is el az egyes lépéseket!
"Határozzuk meg, hogy a 400m kerületű téglalapok közül melyik területe maximális!"

thought a következőhöz: pár másodperc ▼

endben, nézzük meg lépésről lépésre, hogyan határozzuk meg azt a téglalapot, amelynek kerülete 400 méter és területe maximális!

A probléma megértése

leg kell találnunk azt a téglalapot, amelynek a kerülete 400 méter, és ezen téglalapok közül melyik rendelkezik a legnagyobb területtel.

1. Változók bevezetése

elöljük:

- a a téglalap egyik oldalának hossza,
- b a másik oldal hosszával.

2. A kerület képlete

téglalap kerülete a két oldal hosszának kétszereseinek összege: $2a + 2b = 400$

ebből egyszerűsíthetjük: $a + b = 200, b = 200 - a$

3. A terület képlete

téglalap területe: $T = a \times b$

ivel $b = 200 - a$, behelyettesítve: $T = a \times (200 - a), T = 200a - a^2$

5. A terület maximalizálása

Most meg kell találnunk azt az a értéket, amelynél T a lehető legnagyobb.

Ez egy másodfokú függvény, amit átírhatunk: $T = -a^2 + 200a$

A kvadrátikus függvény maximuma a parabolának a csúcspontjában van. A csúcspont $a = -\frac{b}{2a}$, ahol a másodfokú tag együtthatója itt -1 , és az elsőfokú tag együtthatója 200 .

$$a = -\frac{200}{2 \times (-1)} \quad a = \frac{200}{2} \quad a = 100$$

6. A másik oldal meghatározása

Mivel $b = 200 - a$: $b = 200 - 100 \quad b = 100$

7. Eredmény

A maximális területű téglalap esetén mindkét oldal hossza 100 méter. Ez azt jelenti, hogy a téglalap valójában egy négyzet.

8. Összefoglalás

A 400 méter kerületű téglalapok közül a terület maximális akkor, amikor a téglalap minden oldala egyenlő, azaz négyzet alakú. Ebben az esetben minden oldal hossza 100 méter, és a terület: $T = 100 \times 100 = 10\,000 \text{ m}^2$

1. példa. Egy feladat ChatGPT-vel való megoldása. A modell nemcsak korrektül megoldja a feladatot, de el is magyarázza a megoldást magyar nyelven.

2. Számolja ki az alábbi határértéket.

$$\lim_{x \rightarrow 1} \frac{x^2 - 1}{\sqrt{x} - \sqrt{2-x}}$$

The problem asks to calculate the following limit:

$$\lim_{x \rightarrow 1} \frac{x^2 - 1}{\sqrt{x} - \sqrt{2-x}}$$

Let's proceed with the solution:

1. **Simplify the Numerator:** The numerator $x^2 - 1$ can be factored as:

$$x^2 - 1 = (x - 1)(x + 1)$$

So, the expression becomes:

$$\lim_{x \rightarrow 1} \frac{(x - 1)(x + 1)}{\sqrt{x} - \sqrt{2-x}}$$

2. **Rationalize the Denominator:** To handle the denominator $\sqrt{x} - \sqrt{2-x}$, we can multiply both the numerator and the denominator by the conjugate of the denominator, $\sqrt{x} + \sqrt{2-x}$:

$$\lim_{x \rightarrow 1} \frac{(x - 1)(x + 1)(\sqrt{x} + \sqrt{2-x})}{(\sqrt{x} - \sqrt{2-x})(\sqrt{x} + \sqrt{2-x})}$$

3. **Simplify the Denominator:** Using the difference of squares formula, we have:

$$(\sqrt{x} - \sqrt{2-x})(\sqrt{x} + \sqrt{2-x}) = x - (2-x) = 2x - 2 = 2(x - 1)$$

Substituting this back, we get:

$$\lim_{x \rightarrow 1} \frac{(x - 1)(x + 1)(\sqrt{x} + \sqrt{2-x})}{2(x - 1)}$$

4. **Cancel the Common Factor:** The $(x - 1)$ term cancels out:

$$\lim_{x \rightarrow 1} \frac{(x + 1)(\sqrt{x} + \sqrt{2-x})}{2}$$

5. **Evaluate the Limit:** Substitute $x = 1$ into the remaining expression:

$$= \frac{(1 + 1)(\sqrt{1} + \sqrt{2-1})}{2} = \frac{2 \cdot (1 + 1)}{2} = \frac{2 \cdot 2}{2} = 2$$

The final answer is:

2

2. példa. Példa egy feladat ChatGPT-vel való megoldására screenshot alapján. Látható hogy a modell egy egyetemi zh feladatot is képes egyszerűen megoldani.

A. 861. Legyen $f(x) = x^2 - 2$. Jelölje $f^{(n)}(x)$ a függvény n -szeres iteráltját, azaz legyen $f^{(1)}(x) = f(x)$ és $f^{(k+1)}(x) = f(f^{(k)}(x))$.

Legyen $H = \{x: f^{(100)}(x) \leq -1\}$. Határozzuk meg a H halmaz hosszát (a H -t alkotó intervallumok hosszainak összegét). A megoldást zárt alakban várjuk.

Javasolta: *Matolcsi Dávid* (Budapest)

(7 pont)

To solve this problem, we need to determine the total length of the set $H = \{x : f^{(100)}(x) \leq -1\}$, where $f(x) = x^2 - 2$ and $f^{(n)}(x)$ denotes the n-th iterate of f. Step 1: Understanding the Iteration The function $f(x) = x^2 - 2$ is a quadratic function with a graph that opens upwards, and its vertex is at $(0, -2)$. The critical points and fixed points of f are essential in understanding its iterative behavior: Fixed Points: Solve $f(x) = x$: $x^2 - 2 = x \implies x^2 - x - 2 = 0 \implies x = \frac{1 \pm \sqrt{1+8}}{2} = \frac{1 \pm 3}{2}$. So, $x = -1$ and $x = 2$ are fixed points. Behavior Outside $[-2, 2]$: For $x \geq 2$, $f(x) \geq 2$. This means values outside $[-2, 2]$ quickly grow beyond -1 upon iteration. Step 2: Mapping Iterations Using Trigonometric Identities We notice that iterating f can be elegantly described using trigonometric functions: - Let $x = 2 \cos \theta$, where $\theta \in [0, \pi]$. - Then: $f(x) = f(2 \cos \theta) = (2 \cos \theta)^2 - 2 = 4 \cos^2 \theta - 2 = 2 \cos 2\theta$ - By induction, we find: $f^{(n)}(x) = 2 \cos(2^n \theta)$	Step 3: Determining the Condition $f^{(100)}(x) \leq -1$ We need to find all x such that $f^{(100)}(x) \leq -1$: - Transforming the inequality: $f^{(100)}(x) = 2 \cos(2^{100} \theta) \leq -1 \implies \cos(2^{100} \theta) \leq -\frac{1}{2}$ - The solutions for $\cos \phi \leq -\frac{1}{2}$ are in the intervals: $\phi \in [\frac{2\pi}{3}, \pi] \cup [-\pi, -\frac{2\pi}{3}]$ - Since $2^{100} \theta$ wraps around the unit circle 2^{100} times, the proportion of θ satisfying the condition remains constant due to the periodicity and uniform distribution of angles. Step 4: Calculating the Measure of H - The proportion of θ in $[0, \pi]$ satisfying $\cos \phi \leq -\frac{1}{2}$ is $\frac{1}{3}$. - Since $x = 2 \cos \theta$ maps $\theta \in [0, \pi]$ to $x \in [-2, 2]$, the total length of $[-2, 2]$ is 4. - Therefore, the measure of H is: Length of $H = \frac{1}{3} \times 4 = \frac{4}{3}$ Answer: $\frac{4}{3}$
--	---

3. példa. Példa egy versenyfeladat ChatGPT-vel való megoldására. Ugyan a megoldás nem tökéletes, de tartalmazza a lényeges észrevételeket, fő ötleteket, melyek segítségével eljut a helyes megoldásra.

3.3. ChatGPT magyar és angol nyelven

Érdekes megvizsgálni azt is, hogy egy nagy nyelvi modell esetében számít-e a feladat nyelve, hiszen elviekben képes lehet önállóan lefordítani a kapott szöveget. Korábbi interakcióim alapján azt tapasztaltam, hogy a modell magyarul némileg pontatlanabb és néha kissé hamiskásan fogalmaz, de azt feltételeztem, hogy a matematika területén ez a különbség kevésbé lesz jelentős, mivel a számok világa nyelvfüggetlen. Ennek vizsgálatára leteszteltem a modell teljesítményét a 2024-es emelt szintű érettségi matematika feladatain, először magyarul, majd ugyanazokat a feladatokat angolra fordítva is megoldattam vele egy új ablakban. Az eredmény: a teljesítmény szinte teljesen azonosnak bizonyult. A magyar feladatokon mindössze 1 ponttal ért el kevesebbet, mint az angolon (lásd 5. ábra).

Feladatok sorszáma	Maximum pontszám	4o magyar	4o angol	Abszolút különbség
1a	6	6	6	0
1b	5	3	4	1
1c	3	3	3	0
2a	4	4	1	3
2b	4	4	4	0
2c	3	3	3	0
3a	5	2	2	0
3b	3	3	1	2
3c	4	4	4	0
4a	4	1	4	3
4b	6	4	3	1
4c	4	4	4	0
5a	3	1	1	0
5b	3	3	3	0
5c	5	3	5	2
5d	5	0	2	2
6a	5	1	2	1
6b	9	8	8	0
6c	2	1	2	1
7a	5	5	5	0
7b	3	3	3	0
7c	8	2	3	1
8a	3	3	2	1
8b	8	0	0	0
8c	5	2	1	1
9a	5	5	5	0
9b	5	5	3	2
9c	6	3	3	0
Összesen:	131	86	87	

5. ábra. A GPT-4o által elért részeredmények a magyar és angol feladatlapon. Érdekességgént megjegyezhető, hogy a hibák is túlnyomórészt ugyanazokban a feladatokban jelentkeztek, ezt az abszolút különbség oszlopból látni.

3.4. Matematika Érettségi

Az első tesztfeladatsor, melyen kiértékeltem a ChatGPT teljesítményét az a matematika érettségi, melyet minden továbbtanulni vágyó hallgatónak kötelezően el kell végeznie középiskolai tanulmányai végén.

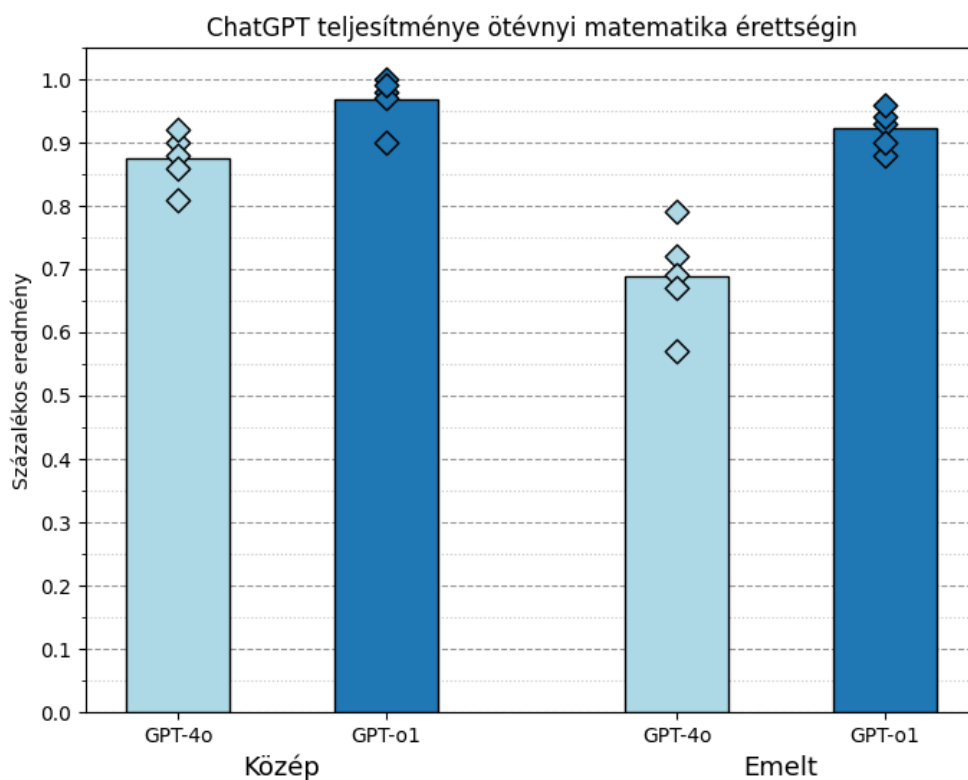
A tanulók matematika érettségét kétféle különböző szinten tehetnek, közép, illetve emelt szinten. A középszintű érettségi alapvetően két részből áll: a dolgozat első felén rövidválaszos egyszerű feladatokat kell megoldaniuk (1-12. feladatok), melyekre 2-3 pontokat kaphatnak, majd a második felében kifejtős számolós feladatokat kell megoldaniuk (13-18. feladatok), melyek összetettebbek, több feladatrészből állnak, és részletesebb indoklást is várnak el a javítás során. A második blokk utolsó három feladatából csak kettőt kell megoldaniuk a hallgatóknak. Emelt szinten egy részből áll az érettségi feladat, kilenc összetett kifejtős feladatot kell megoldaniuk az érettségizőknek, melyben az utolsó 5 feladatból elegendő 4-et megoldani.

Ebben a fejezetben 5 évnyi érettségi feladatsort értékeltem ki, mind közép, illetve emelt szinten, összesen több, mint 270 feladatot. A feladatokat, mind a GPT-4o, és GPT-o1 modellekkel megoldattam, és ezeket a válaszokat vizsgáltam. Egyesével végigolvastam a modellek által generált megoldásokat és gondolatmeneteket, majd ezeket manuálisan, a hivatalos megoldókulcs alapján pontoztam. Ezeket excel táblába gyűjtöttem (lásd 6. ábra).

év, %%	ex num	answer ori	answer key	type	max pont	elért	végleges eredm	hiba 1	hiba 2
	2b	2049 integer solutions	2020 integer solutions	algebra	6	3	nem	nem vesz figyelembe	
	3a	260,000 people seeking jobs	260,000 people	algebra	3	3	igen	-	
	3b	67.85% of the working-age population	0,68	algebra	4	4	igen	-	
	3c	Numbers: {1, 2, 3, 4, 5, 100, 200}, Mean = 45, Median = 4	Numbers: {1, 2, 3, 4, 5, 6, 21}, Mean = 6, Median = 4	statistics	3	3	igen	-	
	3d	Mean can be skewed by outliers; conclusion is incorrect	Correct	statistics	2	2	igen	-	
	4a	Center: (2.8, 5.6), Area: 78.32 cm ²	Center: (2.8, 5.6), Area: 78.4 cm ²	geometry	7	7	igen	-	
	4b	21 cm ²	9 cm ²	calculus	7	3	nem	rossz számmal számolt	
	5a	126 ways	15 ways	combinatorics	5	2	nem	hibás út	
	5b	Probability = 14/75	Probability = 1/60	probability	5	2	nem	félre értelmez	

6. ábra. Az excelben az elért és maximum pontokon túl arról is gyűjtöttem adatot, hogy melyik témakörbe tartozik a feladat, mik voltak a végeredményei (valamint, hogy azok egyeznek-e) és a hibák típusáról is.

Összesítve az eredményeket az 7. ábrán látható teljesítményt figyelhetjük meg. Elmondható, hogy középszinten, mind a GPT-4o, mind a GPT-o1 modell jól teljesítettek, rendre 88% és 96%-os átlagteljesítménnyel. Emelt szinten viszont a GPT-4o teljesítménye lecsökkent 69%-ra, mely egyébként még jeles szintnek felel meg, azonban a GPT-o1 modell emelt szinten is 90% feletti teljesítményt produkált. Évekre lebontva a kiértékelt teljesítményt tartalmazza az 1. és 2. táblázatok. Alapvetően megfigyelhető, hogy a középszintű egyszerű rövidválaszos feladatok esetében mindkét modell nagyon jól teljesít, egy-két félreértéstől eltekintve kiválóan megoldják azokat. A komplexebb kifejtősebb feladatok esetében már középszinten is találhatók jelentősebb eltérések a két modell teljesítménye között, de emelt szinten már még inkább meglátszik ez a különbség.



7. ábra. A ChatGPT teljesítménye az általam kiértékelt 5 évnyi adaton. Összességében elmondható, hogy középszinten mindkét modell elég jó teljesítményt produkált, de emelt szinten a GPT-4o sokkalta gyengébben teljesített a GPT-o1 modellhez képest.

Középszintű matematika érettségi

t 2017			Középszint 2018				Középszint 2019				Középszint 2021				Középszint 2022			
4o	o1	max pont	Feladat	4o	o1	max pont	Feladat	4o	o1	max pont	Feladat	4o	o1	max pont	Feladat	4o	o1	max pont
2	2	2	1	2	2	2	1	2	2	2	1	2	2	2	1	2	2	2
2	2	2	2	2	2	2	2	1	2	2	2	1	2	3	2	2	2	2
2	2	2	3	2	2	2	3	2	2	2	3	2	2	2	3	2	2	2
3	3	3	4	2	2	2	4	4	4	4	4	2	2	2	4	2	2	2
2	2	2	5	2	2	2	5	2	2	2	5	2	2	2	5	2	2	2
2	2	2	6	3	3	3	6	2	2	2	6	2	2	2	6	2	2	2
0	2	2	7	3	3	3	7	2	2	2	7	4	4	4	7	0	2	2
4	4	4	8	2	2	2	8	1	2	2	8	4	4	4	8	2	2	2
2	2	2	9	2	2	2	9	2	2	2	9	2	2	2	9	3	3	3
2	2	2	10	1	3	3	10	4	4	4	10	2	2	2	10	3	3	3
4	4	4	11	4	4	4	11	3	3	3	11	2	2	2	11	2	3	3
3	3	3	12	4	4	4	12	3	3	3	12	3	3	3	12	2	4	4
9	10	10	13	11	11	12	13	10	10	11	13	9	13	13	13	10	10	11
11	12	12	14	8	8	10	14	7	11	12	14	9	11	11	14	12	12	12
9	14	14	15	7	7	14	15	13	13	13	15	12	12	12	15	11	13	13
14	14	17	16	15	17	17	16	17	17	17	16	15	15	17	16	14	17	17
17	17	17	17	17	17	17	17	17	17	17	17	8	8	17	17	9	9	17
16	17	17	18	17	17	17	18	12	12	17	18	13	17	17	18	4	17	17

1. táblázat A ChatGPT teljesítménye középszintű matematika érettségin évekre lebontva. Az egyszerű rövidválaszos feladatok esetében mindkét modell jól teljesít, a GPT-4o néhány félreértéstől eltekintve kiválóan megoldja, a GPT-o1 modell pedig szinte hibátlanra. A kifejtősebb jellegű második blokkban már található nagyobb eltérés a két modell teljesítménye között. Egy-két helyen meglepően alacsony pontszámot tapasztalhatunk o1 esetében is, ennek oka, hogy az o1 modell nem támogat kép formátumú mellékleteket, így az olyan jellegű feladatokat, melyekhez ábra is tartozik, nem tud egyelőre megoldani.

Emelt szintű matematika érettségi

2020			
4o	o1	max pont	
0	12	12	
9	13	13	
12	12	12	
10	14	14	
4	16	16	
8	8	16	
15	15	16	
12	12	16	
11	16	16	

Emeltszint 2021				
Feladat	4o	o1	max pont	
1	7	13	13	
2	5	5	12	
3	8	8	13	
4	8	13	13	
5	16	16	16	
6	6	16	16	
7	9	16	16	
8	0	16	16	
9	6	6	16	

Emeltszint 2022				
Feladat	4o	o1	max pont	
1	6	11	14	
2	10	10	10	
3	1	13	13	
4	13	13	14	
5	12	12	16	
6	11	11	16	
7	6	16	16	
8	13	13	16	
9	13	13	16	

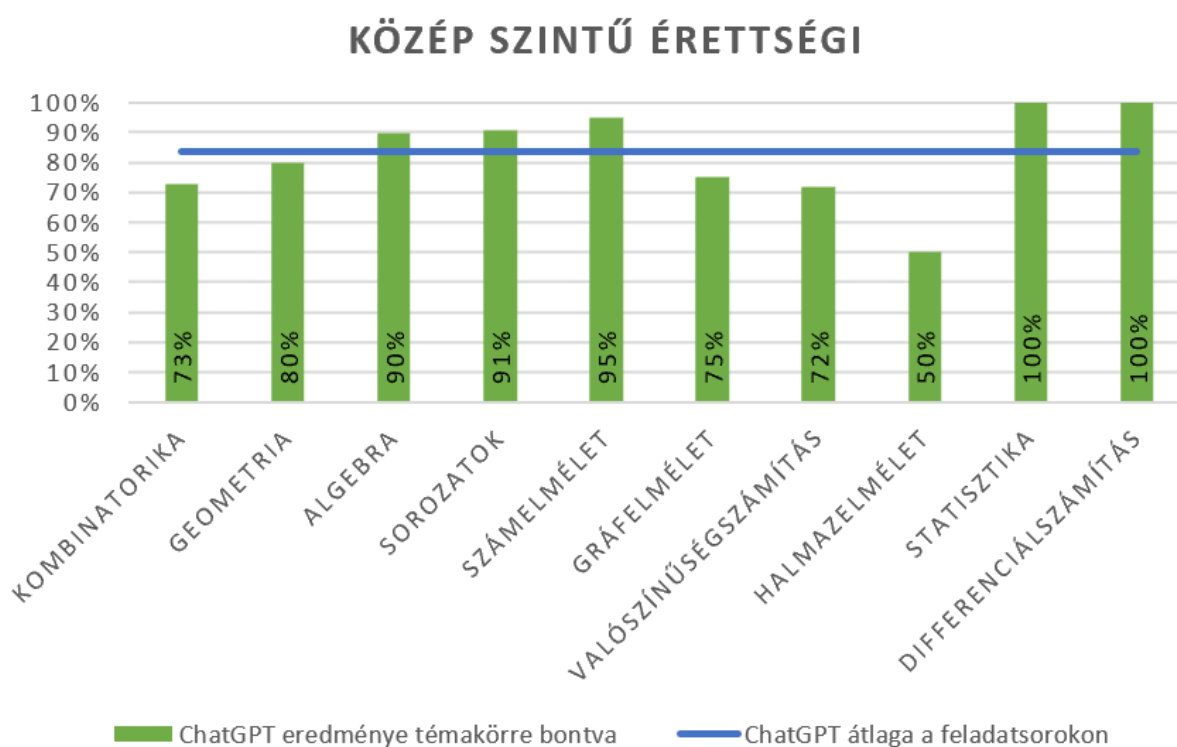
Emeltszint 2023				
Feladat	4o	o1	max pont	
1	10	10	11	
2	8	8	14	
3	14	14	14	
4	8	12	12	
5	2	16	16	
6	12	16	16	
7	16	16	16	
8	7	7	16	
9	16	16	16	

Emeltszint 2024				
Feladat	4o	o1	max pont	
1	8	8	13	
2	11	11	13	
3	13	13	13	
4	10	12	12	
5	3	12	16	
6	5	16	16	
7	5	16	16	
8	15	15	16	
9	16	16	16	

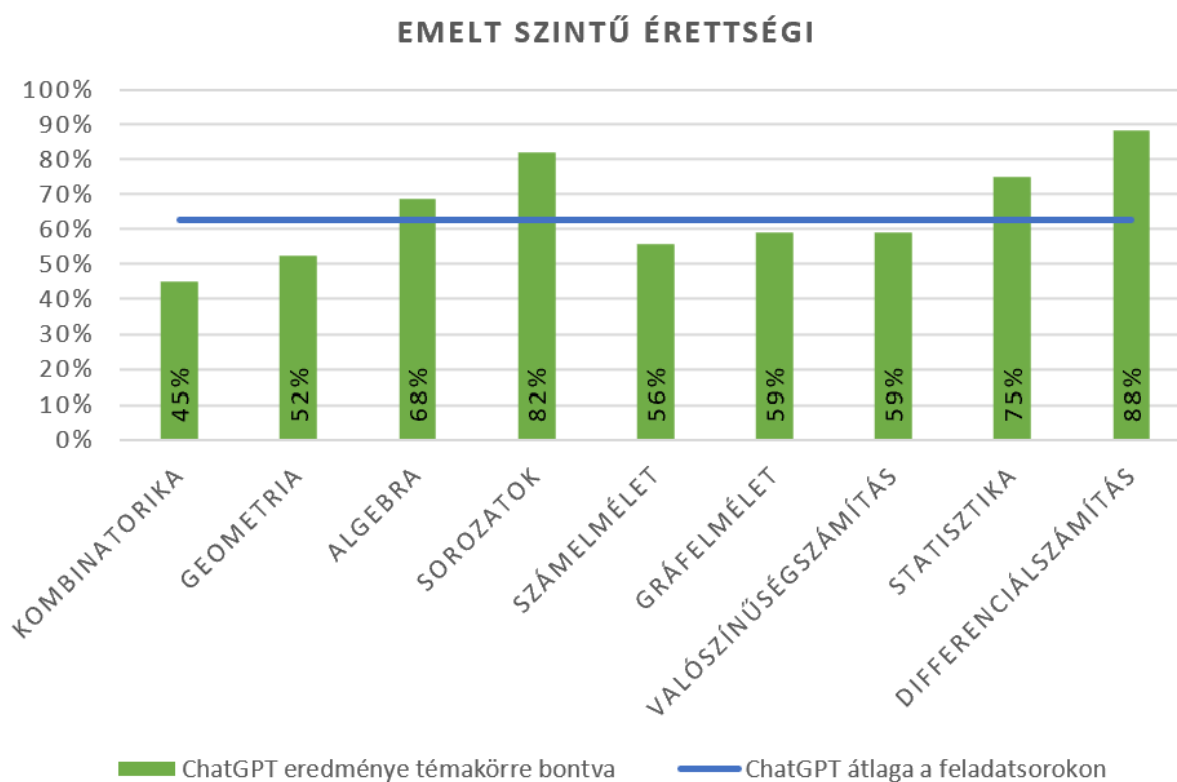
2. táblázat A ChatGPT teljesítménye emelt szintű matematika érettségien évekre lebontva. A középszintű érettségikkel ellentétben itt már megfigyelhető minden téren komolyabb eltérés a 4o és o1 modell teljesítményében, például több feladat esetében is legalább 10 pontnyi javulás érhető el a jobb modell használatával.

3.5. Eredmény témakörönként

A feladatokat témakörök szerint is csoportosítottam, így lehetőség nyílik annak vizsgálatára, hogy az egyes témakörökben hogyan teljesít a modell (lásd 8. és 9. ábra). Összességében megfigyelhető, hogy a feladatok sikeres megoldását befolyásolja, hogy melyik témakörből származnak a kérdések, mivel mind középszinten, mind emelt szinten vannak olyan feladatok, amelyek a modell átlagon felül vagy éppen átlag alatti teljesítményt nyújt. A sorozatok és a differenciálszámítás témakörébe tartozó feladatokat például kiválóan oldja meg, míg a kombinatorika egyértelműen a gyenge pontjai közé tartozik.



8. ábra. A 4o modell teljesítménye középszintű érettségien témakörönként

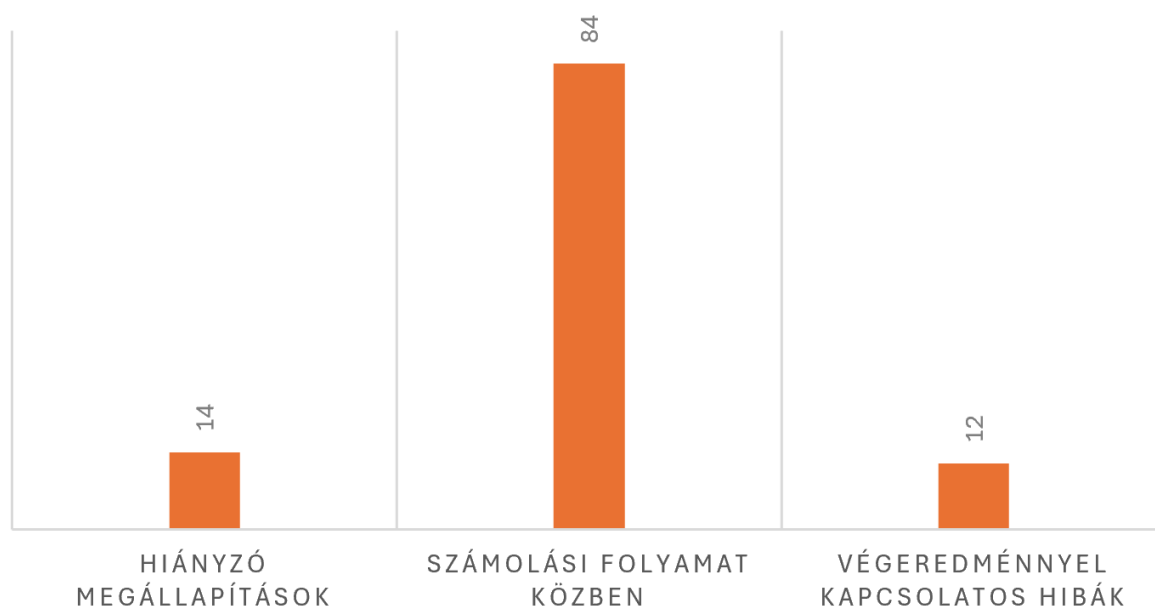


9. ábra. A 4o modell teljesítménye emelt szintű érettségien témakörönként

3.6. Tipikus hibák

A kézi ellenőrzés során, azoknál a feladatoknál, ahol a modell nem érte el a maximum pontszámot, részletesen megvizsgáltam a hibák természetét, és kategóriákba soroltam őket (lásd 10. ábra). Összesen három fő csoportot hoztam létre, amelyekben további alkategóriákat alakítottam ki, hogy jobban át tudjam tekinteni a hibák típusait. A következőekben bemutatom a különböző kategóriákat és hozzájuk tartozó megfelelő példákat.

HIBÁK KATEGÓRIÁNKÉNT



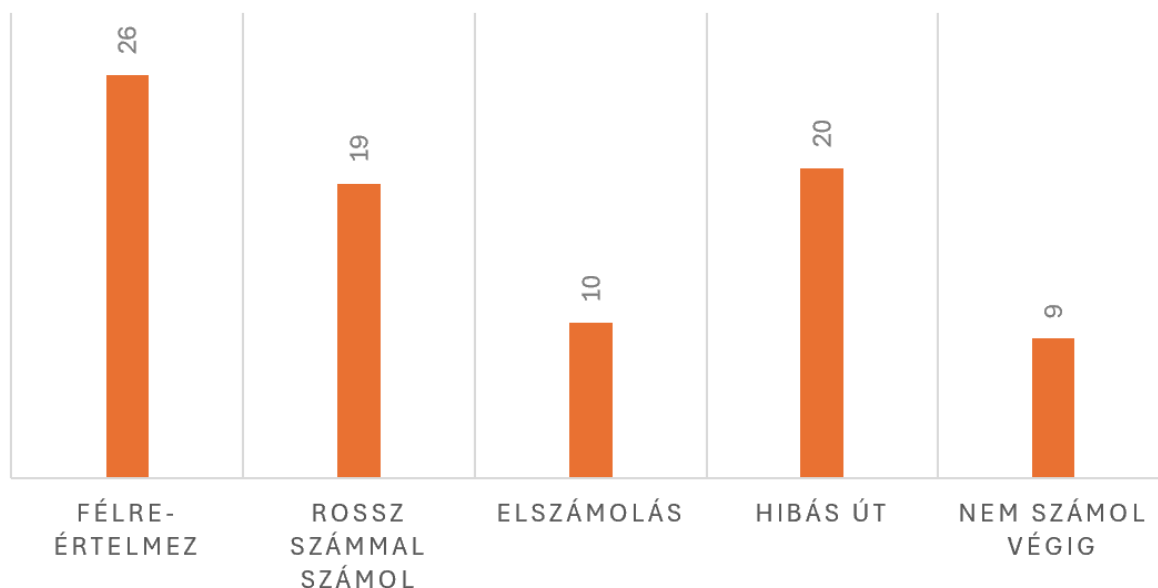
10. ábra. Az összes hiba kategóriák szerinti eloszlása. Nem meglepő, hogy a legtöbb hiba a második kategóriában jelentkezett, mivel ebben a fázisban voltak a legösszetettebb lépések. Fontos még megjegyezni, hogy az első és harmadik kategóriás hibákat főleg a 4o modell követte el, az alaposabb o1 modell ilyen hibákat nem vétett.

A kategóriák a következők voltak:

1. Hiányzó megállapítások (“figyelmetlenségi” hibák): Ebbe a kategóriába főként olyan hibák tartoztak, melyet diákok is sokszor elkövetnek, de csak kisebb pontlevonással járnak. Ezek például olyan esetek, mint amikor a modell nem ellenőrizte a végeredményt olyan feladatoknál, ahol erre szükség lett volna, vagy trigonometrikus számításoknál nem jelezte, hogy ekvivalens átalakításokat végzett. Ezek az apró hibák sokszor nem befolyásolták a végeredményt, amely egyébként helyes volt.

2. Számolási folyamat közben felmerülő hibák: Ez a kategória foglalta magában a legtöbb hibát, ezért külön altípusokat különböztettem meg (lásd 11. ábra):

2. KATEGÓRIA LEBONTVA



11. ábra. 2. Kategóriás hibák lebontva altípusokra

o Félreértelmezés: Ebbe azok a hibák tartoznak, amikor a modell rosszul értelmezte a probléma egy kulcsfontosságú elemét (lásd 12. ábra), például egy korlátozást (lásd 13. ábra), ami helytelen egyenlet felírásához vagy hibás gondolatmenethez vezetett. Gyakori volt az is, hogy nem vette figyelembe az egyes feladatokban szereplő kikötéseket, vagy helytelenül olvasta be az adattáblákban szereplő információkat.

3. a) Számítsa ki annak a valószínűségét, hogy egy játékos első dobása Sor lesz!

Egy "Sor" számkombináció esetén az öt kockán öt különböző, egymást követő szám szerepel. Lehetséges sorozatok a dobókockával:

- 1, 2, 3, 4, 5
- 2, 3, 4, 5, 6

Tehát összesen kétféle "Sor" számkombináció létezik az ötdobásos lehetőségek közül.

Mivel minden kockának hat oldala van, összesen $6^5 = 7776$ különböző eredményt dobhatunk öt kockával.

A két lehetséges "Sor" kombináció mindegyike pontosan 1-féleképpen alakulhat ki (pl. 1, 2, 3, 4, 5 és 2, 3, 4, 5, 6), így a "Sor" számkombinációk száma mindössze 2.

A valószínűség kiszámítása:

$$\frac{2}{7776} = \frac{1}{3888} \approx 0,000257$$

Tehát annak a valószínűsége, hogy egy játékos első dobása Sor lesz, körülbelül 0,000257 (0,0257%).



3. a) második megoldás	
A Sor számkombináció 2-féle lehet (1-2-3-4-5 vagy 2-3-4-5-6).	1 pont
Feltehetjük, hogy az öt kockával egyesével dobunk. Ekkor annak a valószínűsége, hogy a játékos 1-2-3-4-5 számkombinációt dob valamilyen sorrendben	2 pont
$\frac{5}{6} \cdot \frac{4}{6} \cdot \frac{3}{6} \cdot \frac{2}{6} \cdot \frac{1}{6} = \frac{120}{7776}$ (Elsőre 5-félét, másodikkra 4-félét dobhatunk, stb.)	
Ugyanennyi annak a valószínűsége, hogy a játékos 2-3-4-5-6 számkombinációt dob valamilyen sorrendben.	1 pont
A kért valószínűség tehát	1 pont
$\frac{240}{7776} \left(= \frac{5}{162} \right) \approx 0,031.$	
Összesen:	5 pont

12. ábra. Példa félreértelmezésre. Itt a modell úgy veszi, mintha a sort csak helyes sorrendben lehetne kidobni, de igazából bármilyen sorrendben ér.

We need to find the number of seven-digit binary numbers with at most two 0 digits. This includes cases where there are:

1. Zero 0s (all 1s).
2. One 0.
3. Two 0s.

Let's go through each case:

1. **Zero 0s (all 1s):**

There's only one such number: `1111111`.

2. **One 0:**

We can place this single 0 in any of the seven positions, so there are $\binom{7}{1} = 7$ numbers.

3. **Two 0s:**

We can place two 0s in any two of the seven positions, so there are $\binom{7}{2} = 21$ numbers.

Now, we add up these possibilities:

$$1 + 7 + 21 = 29$$

2. a)	
Egy olyan megfelelő szám van, amely nem tartalmaz 0 számjegyet (1111111).	1 pont
Ha a szám egy 0 számjegyet tartalmaz, akkor az a legnagyobb helyiértéken kívül bárhol lehet, tehát 6 ilyen szám van.	1 pont
Ha a szám két 0 számjegyet tartalmaz, akkor ezek a legnagyobb helyiértéken kívül bárhol lehetnek, így helyüket $\binom{6}{2} = 15$ -féleképpen választhatjuk ki.	1 pont
Összesen tehát $(1 + 6 + 15 =) 22$ megfelelő szám van.	1 pont
Összesen:	4 pont

Megjegyzés: Teljes pontszámot kapjon a vizsgázó, ha rendezetten felmokat, és ez alapján helyesen válaszol.

13. ábra. Még egy példa egy korlátozás figyelmen kívül hagyására: 0 nem szerepelhet elsőként, hiszen akkor csak hatjegyű szám lenne.

o Hibás út: Ide főként elvi hibák tartoznak, amikor egy megoldásra olyan utat választott a modell, amellyel a fókuszát áthelyezte más számolásokra és eredményekre, aminek következtében nem a kérdésre adott választ, vagy rossz választ adott.

o Rossz számmal számolt: Itt azok a hibák szerepelnek, amelyek során a modell helyesen végezte el a műveleteket, de a számok, amelyekkel dolgozott, hibásak voltak. Ez akkor fordult elő, amikor téves feltételezésre alapozott (lásd 16. ábra), és emiatt hibás részeredmények születtek. Ide soroltam azokat a hibákat is, amelyek a mértékegységek helytelen átváltása miatt következtek be, például méréseknél vagy olyan feladatoknál, ahol az egységek konzisztenciája fontos.

7. c) Határozza meg az 1,8 méter alkotójú forgáskúpok közül annak a sugarát és a magasságát, amelynek a térfogata maximális! Mekkora ez a maximális térfogat?

Az alkotó hossza 1,8 méter, tehát a maximális térfogatú kúp akkor keletkezik, ha a kúp belső háromszöge egyenlő szárú, ami akkor lehetséges, ha a kúp magassága és sugara egyenlőek. Így:

$$h = r$$

A Pitagorasz-tétel alapján:

$$r^2 + h^2 = 1,8^2$$

Mivel $r = h$, ez lesz:

$$2r^2 = 1,8^2$$

$$r^2 = \frac{1,8^2}{2} = \frac{3,24}{2} = 1,62$$

$$r = h = \sqrt{1,62} \approx 1,272 \text{ m}$$

A maximális térfogat:

$$V_{\max} = \frac{1}{3} \pi r^2 h = \frac{1}{3} \pi (1,272)^2 \times 1,272$$

$$V_{\max} \approx \frac{1}{3} \pi \times 1,618 \times 1,272 \approx 0,6867 \pi \approx 2,158 \text{ m}^3$$

Tehát a maximális térfogatú kúp sugara és magassága egyaránt körülbelül 1,272 m, a maximális térfogata pedig körülbelül 2,158 m³.

7. c) első megoldás	
Jelölje a kúp alapkörének sugarát r , magasságát m , ekkor $r^2 + m^2 = 1,8^2$, ahonnan $r^2 = 3,24 - m^2$.	1 pont
A forgáskúp térfogata: $\frac{r^2 \pi m}{3} = \frac{(3,24 - m^2) \cdot \pi m}{3} = \frac{\pi}{3} \cdot (3,24m - m^3)$.	1 pont
A $V:]0; 1,8[\rightarrow \mathbf{R}; V(m) = \frac{\pi}{3} \cdot (3,24m - m^3)$ függvénynek ott lehet maximuma, ahol a deriváltja nulla.	1 pont
$V'(m) = \frac{\pi}{3} \cdot (3,24 - 3m^2) = 0$	1 pont
Innen ($m > 0$ miatt) $m = \sqrt{1,08} \approx 1,04 \text{ m}$.	1 pont
Az $m = \sqrt{1,08}$ helyen a deriváltfüggvény pozitívából negatívba megy át, ezért itt V -nek (abszolút) maximuma van.	1 pont
Ekkor $r (= \sqrt{3,24 - 1,08} = \sqrt{2,16}) \approx 1,47 \text{ m}$,	1 pont
a maximális térfogat pedig kb. $2,35 \text{ m}^3$.	1 pont
Összesen:	8 pont

14. ábra. Példa egy hibás feltételezésre, ami miatt rossz eredmény jön ki.

o Szimpla elszámolás: Ide tartoznak azok az esetek, amikor az algebrai műveleteket helytelenül végezte el (lásd 14. és 15. ábra), vagy kerekítési hibák adódtak, amelyek az eredmény nagy mértékű pontatlanságát okozták.

$$\binom{90}{5} = \frac{90 \times 89 \times 88 \times 87 \times 86}{5 \times 4 \times 3 \times 2 \times 1} = 43,949,268$$

3. The number of possible combinations where all numbers are at least 5 is:

$$\binom{86}{5} = \frac{86 \times 85 \times 84 \times 83 \times 82}{5 \times 4 \times 3 \times 2 \times 1} = 32,961,728$$

4. The probability that Panni gets her wish is the ratio of favorable outcomes to total outcomes:

$$P(\text{Panni gets her wish}) = \frac{32,961,728}{43,949,268} \approx 0.75$$

So, the probability that Panni will get her wish is **approximately 0.75** or **75%**.

15. ábra. Példa az elszámolásra, $\binom{86}{5}$ valójában 34826302.

$$P(X = 1) = \binom{10}{1} (0.08)^1 (0.92)^9 = 10 \times 0.08 \times (0.92)^9 \approx 10 \times 0.08 \times 0.5132 \approx 0.4106$$

16. ábra. Még egy példa elszámolásra, $0,92^9$ valójában 0,4722.

o Nem számol végig: Olyan is előfordult, hogy a modell elkezdte a feladat megoldását, rész-eredményeket helyesen kiszámolt, de amikor egy összetettebb lépéshez érkezett, nem számolta végig a megoldást. Ilyenkor leírta, hogy mely számokat kellene egy adott képletbe beírni, de ő maga nem számította ki a végső eredményt.

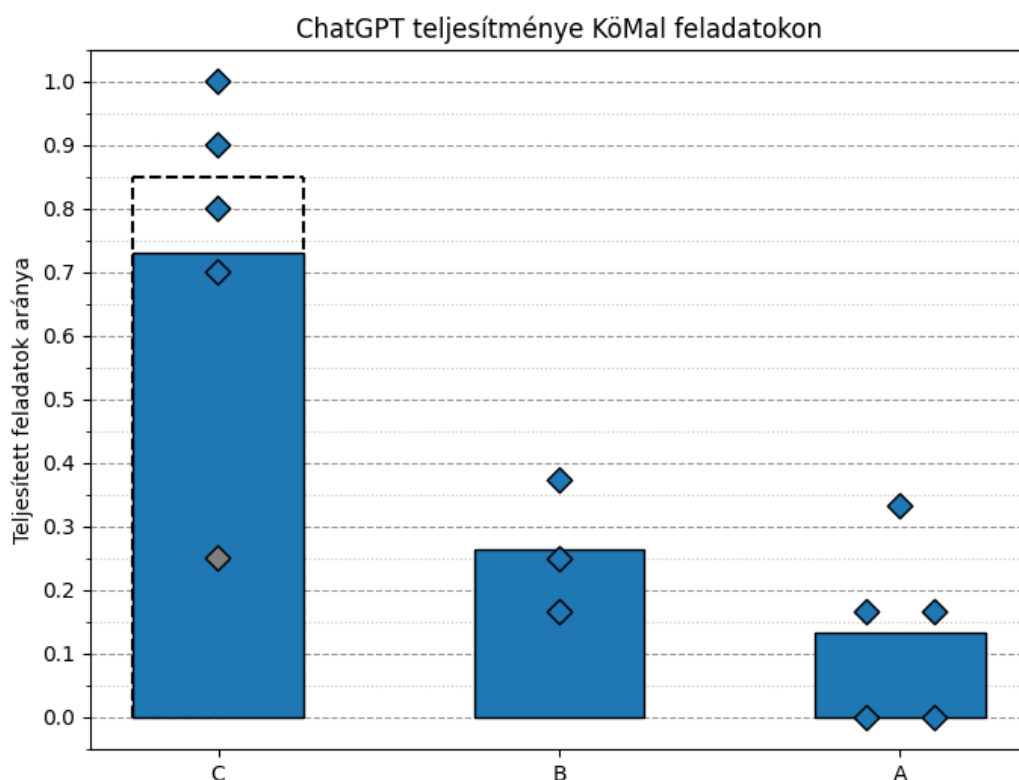
3. Végeredménnyel kapcsolatos hibák: Ide soroltam azokat a hibákat, amikor a modell nem adott meg minden helyes választ vagy rosszul adta meg a végeredményt. Például a helyes számolások ellenére végeredményként egy másik számot írt be, vagy egy másodfokú egyenletnél nem adta meg az összes lehetséges helyes megoldást.

3.7. KöMaL

A Középiskolai Matematikai és Fizikai Lapok (KöMaL) minden évben meghirdeti pontversenyeit, melyben különböző matematika, fizikai és informatikai feladatokat kell megoldaniuk a versenyző hallgatóknak. Ebben a dolgozatban a havonta kiírásra kerülő matematika feladatso-
rok nagy nyelvi modell általi megoldhatóságának elemzésével foglalkoztam. A KöMaL feladatait különböző kategóriákba sorolva hirdeti meg minden hónapban: K, C, B és A kategóriákban. A K-jelű feladatok könnyebbek, melyeket kifejezetten kilencedik osztályosok számára hirdetnek meg, míg a C jelű feladatok vegyes évfolyamnak szólnak, és olyan feladatokat tartalmaznak, melyek nagy mértékben kötődnek az iskolai tanulmányokhoz, és nem rugaszkodnak el túlságosan nehézségben a B és A kategóriákhoz képest. Az A és B jelű versenyfeladatok viszont már nehezebb kategóriába tartoznak, ezek igazi versenyfeladatok, melyek közül az A jelűek már nemzetközi matematikaversenyen való megmérettetésre is felkészítik a hallgatókat. Megjegyzendő, hogy ugyan a C kategória szorosan kapcsolódik az iskolai tananyaghoz, azért azok nehezebbek lehetnek az emelt szintű érettségi feladatainál, mivel figyelembe kell venni, hogy KöMaL esetében az emelt szintű érettségivel ellentétben nem 35 perc jut egy feladatra, hanem egy hónapon keresztül lehet gondolkozni rajtuk, akár 2-3 fős csapatokban.

Mivel korábban már tapasztaltuk, hogy a GPT-4o modellnek meg-meggyűlik a baja az emelt szintű érettségi feladataival, míg a GPT-o1 modell pedig kiválóan teljesít rajtuk, ezért úgy döntöttem, hogy elemzésem során a GPT-o1 modellel foglalkozom, és C-, B- és A-jelű feladatokon értékelem ki. C-jelű feladatok esetén csak azokkal foglalkozom melyek nem K/C jelűek, azaz az idősebb osztályok számára szántakkal foglalkozom.

Versenyfeladatok esetében az érettségivel ellentétben nem feltétlenül arra helyeztem a hangsúlyt, hogy meg tudja-e oldani teljes precizitással a feladatot a ChatGPT-o1 vagy sem. Abban az esetben, ha valami apró hibát vétett, például számolási hibát, mely nem befolyásolja nagyban a feladat megoldását, vagy a megoldás minden fő ötletét tartalmazta, akkor teljesítettnek vettem a feladatot, míg abban az esetben, ha valami lényegi lépést el tudott érni a feladat megoldása során, de nem teljeskörűen oldotta meg, például hagyott egy-egy esetet (de az általa bizonyított eset sem triviális), akkor fél pontot adtam a megoldásra. Megjegyzendő, hogy nagyon ritkán került sor fél pont adására a feladatok megoldása során, általában vagy nagyon szépen megoldotta a feladatot, vagy teljesen rosszul teljesített.



17. ábra. A ChatGPT teljesítménye különböző jelű KöMaL feladatokon. Megfigyelhető, hogy egy feladatsortól (szürke rombusz) eltekintve a C-jelű feladatok esetében eléggé jól teljesített a modell, azonban B-jelű feladatok esetén jelentős visszaesés tapasztalható. A-jelű feladatokkal többségében már nem tud mit kezdeni a ChatGPT-o1, mindössze egy feladat esetén tudott felmutatni egy értékelhető megoldást.

Az elemzésben 5 hónapnyi C-jelű (5x5 feladat), 3 hónapnyi B-jelű (3x8 feladat) és 5 hónapnyi C-jelű feladatot (3x5 feladat) vizsgáltam. Az eredmények (17. ábra) alapján elmondható, hogy alapesetben a C jelű feladatokon egészen jól teljesített a ChatGPT-o1, egy feladatsortól eltekintve 4 és fél feladat megoldására képes, ámbár az az egy feladatsor így lehúzta az átlagot. Azonban a B-jelű feladatok esetében már látványos visszaesés tapasztalható, itt már elég kevés feladatra tudott értékelhető megoldást felmutatni a modell, mindössze a feladatok 25%-t tudta megoldani átlagban. Az A jelű feladatok megoldásában pedig azt tapasztaltam, hogy egy-két értékelhető ötlettől eltekintve, mindössze egyetlen feladat megoldására volt képes. Így bátran kijelenthetjük, hogy a GPT-o1 modell az A-jelű feladatok megoldására már biztosan nem képes, és a B-jelű feladatok esetében is elég gyenge teljesítmény tud felmutatni, ezt úgy értelmezhetjük, hogy valahol a C- és B-jelű feladatok között érhetjük el a ChatGPT határait. Tipikusan megfigyelhető volt a feladatok megoldásai során, hogy geometriai problémák esetében szinte mindig koordináta-geometriai módszerekkel akarja megoldani a feladatok a ChatGPT,

mely bizonyos esetekben eléggé elbonyolította a feladatot. Illetve tipikus hibának tekinthető, hogy bizonyítások esetében nagyon gyakran csak egy-egy speciális példán nézi meg az állítás helyességét, és azáltal véli megoldottnak a feladatot.

A dolgozatban nem szemléltetem, de mindenképpen fontosnak találom megjegyezni, hogy nagy nyelvi modellekhez kapcsolódó kutatásokban gyakran alkalmaznak olyan kiértékelési eljárást, melyben a modell segítségével többször is megoldást generálnak az adott feladatra, majd úgy értékelik ki, hogy megnézik van-e jó megoldás a megoldások között. Erre a technikára ebben az elemzésben nem tértem ki, de elképzelhetőnek tartom, hogy egy adott problémára többször, akár 10-szer is legenerálva a választ találkozni fogunk olyan esettel, amelyben valamilyen hasznos ötlet is szerepel. Azonban a chatbotok ily módon történő alkalmazásához már alaposabb matematikai tudásra van szükség, hiszen a felhasználónak ki kell tudnia szűrnie potenciálisan például 9 teljesen rossz válaszból és 1 félig-meddig jó válaszból, hogy mik azok az ötletek, melyek valóban értékesek. Ez persze extra energiabefektetéssel jár, és kritikai látásmódot vár el a felhasználótól, azonban egy komoly matematika verseny esetén egy erős versenyző a chatbotok ily módon történő használatából is profitálhat. A következőekben szemléltetem Dr. Hujter Mihály professzor úr két B-jelű feladatára ChatGPT-o1 megoldását (lásd 4. és 5. példa).

egyenlőség fennáll az n alapú számrendszerben, ha $n \geq 9$.

Javasolta: *Hujter Mihály* (Budapest)

$$\sqrt{\frac{123456787654321}{1234321}} = 10001$$

and result: 10001 in base n .

$$\begin{aligned} & \cdot n^{14} + 2 \cdot n^{13} + 3 \cdot n^{12} + 4 \cdot n^{11} + 5 \cdot n^{10} + 6 \cdot n^9 + 7 \cdot n^8 + 8 \cdot n^7 \\ & 7 \cdot n^6 + 6 \cdot n^5 + 5 \cdot n^4 + 4 \cdot n^3 + 3 \cdot n^2 + 2 \cdot n + 1. \end{aligned}$$

$$D = 1 \cdot n^6 + 2 \cdot n^5 + 3 \cdot n^4 + 4 \cdot n^3 + 3 \cdot n^2 + 2 \cdot n + 1.$$

$$\frac{T}{S} = n^4 + 1.$$

This equality holds in any base $n \geq 9$ because all digits involved are valid digits in base n , and the algebraic manipulations confirm the equality.

4. példa. A trükkös átalakításokkal azonosságok alkalmazásában bővelkedő n alapú számrendszerbeli egyenletes B-jelű KöMal versenyfeladatot a ChatGPT-o1 modell mindössze 52 másodperc alatt megoldja.

B. 5408. Egy háromszög egyik oldala számtani közepe a másik kettőnek. Bizonyítandó, hogy ezen középső oldalt kettévágó szögfelező hossza $\frac{\sqrt{3}}{2}$ -szerese a másik két oldal mértani közepének.

Javasolta: *Hujter Mihály* (Budapest)

Vetkezőhöz: 23 másodperc ▾

in a triangle where one side is the arithmetic mean of the other two sides, the angle bisector from the vertex opposite the longest side has a length equal to $\frac{\sqrt{3}}{2}$ times the geometric mean of the other two sides, step by step.

Let ABC with sides $a = BC$, $b = AC$, and $c = AB$. Since b is the arithmetic mean of sides a and c :

$$b = \frac{a + c}{2}.$$

The angle bisector from vertex B divides side AC into segments AD and DC . We want to find the length of the angle bisector BD is:

$$BD = \frac{\sqrt{3}}{2} \sqrt{ac}.$$

Step 1: Use the Angle Bisector Theorem: The angle bisector from vertex B divides side AC into segments proportional to the adjacent sides:

$$\frac{AD}{DC} = \frac{AB}{BC} = \frac{c}{a}.$$

2. Express AD and DC in Terms of b :

Let $AD = m$ and $DC = n$, so $m + n = b$.

From the Angle Bisector Theorem:

$$\frac{m}{n} = \frac{c}{a} \implies m = \frac{c}{a}n.$$

Since $m + n = b$:

$$\frac{c}{a}n + n = b \implies n \left(1 + \frac{c}{a}\right) = b \implies n = \frac{ab}{a + c}.$$

Similarly:

$$m = \frac{c}{a}n = \frac{c}{a} \cdot \frac{ab}{a + c} = \frac{cb}{a + c}.$$

3. Use the Angle Bisector Length Formula:

The length of the angle bisector BD is given by:

$$BD^2 = ac \left[1 - \left(\frac{b}{a + c}\right)^2\right].$$

Substitute $b = \frac{a+c}{2}$:

$$\frac{b}{a + c} = \frac{a + c}{2(a + c)} = \frac{1}{2}.$$

Therefore:

$$BD^2 = ac \left[1 - \left(\frac{1}{2}\right)^2\right] = ac \left(1 - \frac{1}{4}\right) = ac \left(\frac{3}{4}\right).$$

4. Compute BD :

$$BD = \sqrt{\frac{3}{4}ac} = \frac{\sqrt{3}}{2} \sqrt{ac}.$$

Conclusion: The length of the angle bisector BD is $\frac{\sqrt{3}}{2}$ times the geometric mean of sides a and c .

5. példa. A geometriai B-jelű KöMal versenyfeladatot a ChatGPT-o1 modell mindössze 23 másodperc alatt megoldja a szögfelezőtétel alkalmazásával.

4. Kérdőív

A kutatásom során alkalmazott kérdőív célja, hogy átfogó képet adjon a diákok tapasztalatairól és véleményéről a ChatGPT használatával kapcsolatban. A kérdőív három fő szakaszra tagolódik: általános kérdések és tapasztalatok, a ChatGPT vizsgálata, mint tanár, valamint egy apró kitekintés a jövőbeni szerepeire.

Az első szakaszban olyan alapvető információkat gyűjtök, mint a kitöltők aktuális tanulmányi évfolyama, matematikához való viszonyulásuk, valamint a ChatGPT-s szokásaik. Ezek a kérdések lehetővé teszik, hogy megérthessem a diákok általános hozzáállását és a ChatGPT alkalmazásának jelenlegi trendjeit a matematika oktatásában.

A második szakaszban specifikusabb kérdések találhatók, amelyek a ChatGPT által generált definíciók megértését és felismerhetőségét vizsgálják. Ezek a kérdések arra irányulnak, hogy a diákok mennyire részesítik előnyben a ChatGPT által nyújtott magyarázatokat más definíciókkal szemben, valamint, hogy képesek-e azonosítani, ha egy definíciót a ChatGPT generált.

Ezen túlmenően, a harmadik részben a kérdőív tartalmaz olyan kérdéseket is, amelyek a jövőbeni felhasználási lehetőségeket és a mesterséges intelligencia hatását vizsgálják a diákok kognitív funkcióira és a matematika tanítására. A kérdőív struktúrája és tartalma összhangban áll kutatási céljaimmal és hipotéziseimmel, lehetővé téve, hogy részletesen elemezzem a ChatGPT hatékonyságát és elfogadottságát a matematikai oktatásban.

A kérdőívben szereplő kérdések többféle válaszadási lehetőséget kínálnak, beleértve az egyszerű és többszörös választási opciókat, valamint értékelő skálákat. Ez a sokszínűség biztosítja, hogy a kitöltők pontosan és részletesen kifejtessék véleményüket és tapasztalataikat.

A kérdőívet összesen 107-en töltötték ki középiskolások és egyetemisták vegyesen.

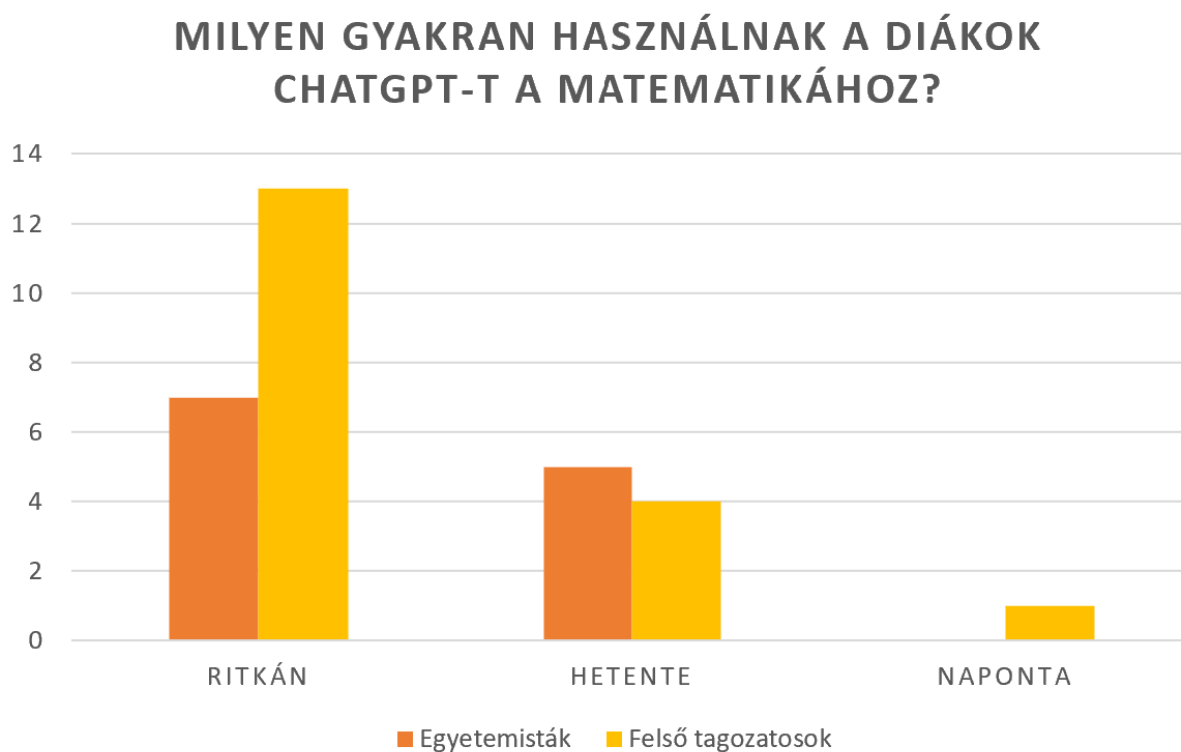
4.1. A kitöltők szokásai, tapasztalatai és véleménye a ChatGPT-vel kapcsolatban

Az első szakaszban általános kérdéseket tettem fel, hogy jobban megismerjem a válaszadókat és később csoportosíthassam őket. Ennek a résznek különlegessége, hogy elágazásokat tartalmaz: a kitöltők a korábbi válaszaik alapján kapják a további kérdéseket. A válaszokból kiderül, hogy összességében az egyetemisták fele már használta a ChatGPT-t, míg a középiskolások körében ez az arány sokkal alacsonyabb, mindössze egynegyedük próbálta ki (lásd 18. ábra).

Évfolyam	Igen	Nem
Egyetem: Alapképzés	42,11%	57,89%
Egyetem: Mesterképzés	57,14%	42,86%
Felső tagozat: 11. - 13. osztály	17,65%	82,35%
Felső tagozat: 7. - 10. osztály	28,13%	71,88%

18. ábra. A diákok eloszlása a ChatGPT használata alapján (táblázat). A 100% 107 kitöltésnek felel meg.

Azok között, akik már használták, jellemző, hogy ritkán (vagy kisebb számban hetente) teszik ezt (lásd 19. ábra). Így kijelenthető, hogy a ChatGPT használata a matematikában egyelőre csak kis mértékben terjedt el.



19. ábra. A diákok eloszlása a ChatGPT használata alapján (diagram).

A diákok matematika iránti viszonyulását is megkérdeztem. Ebben a szakaszban nem találtam korrelációt, mind a négy kategóriában hasonló arányban használták már a ChatGPT-t (lásd 20. ábra).

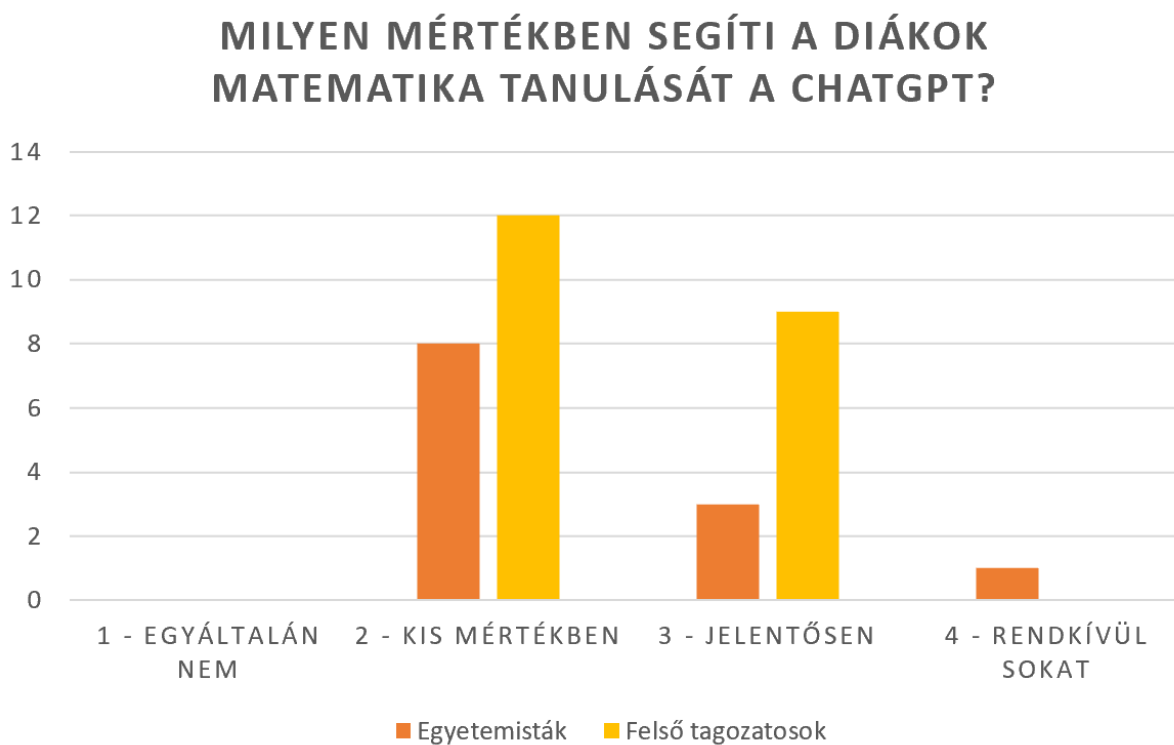
Mennyire szereti a matekot	Igen	Nem
1	42,86%	57,14%
2	22,73%	77,27%
3	31,91%	68,09%
4	31,25%	68,75%

20. ábra. A ChatGPT használata és a diákok matematikához való viszonya

A válaszadók többsége a ChatGPT hasznosságával sem teljesen elégedett, a 4-es skálán az emberek leginkább 2 pontot adtak, páran 3-at, utóbbit főleg a felsőtagozatosok (lásd 21. és 22. ábra).

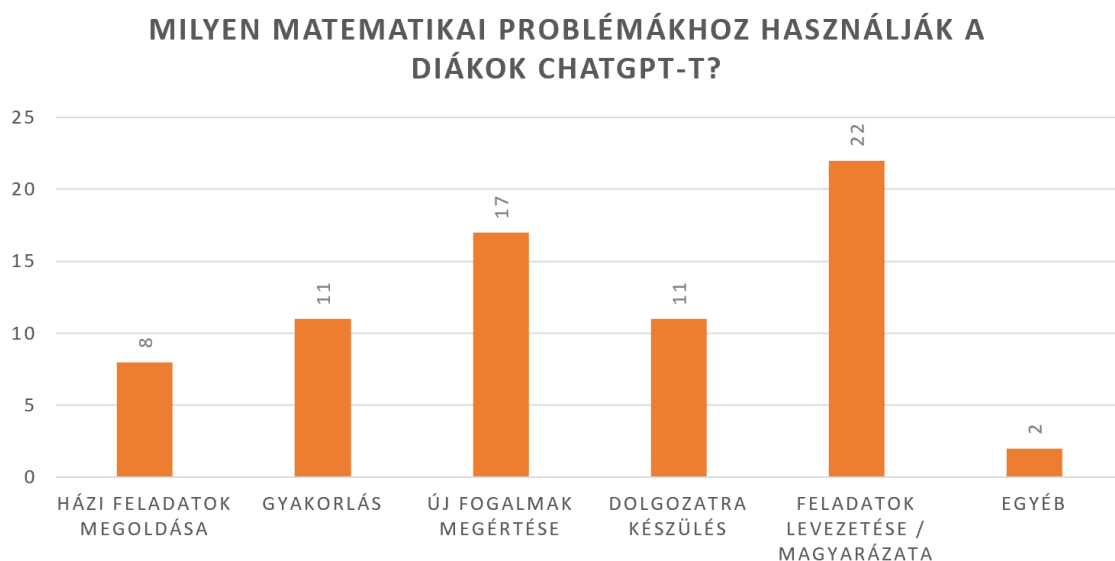
Évfolyam	1 - egyáltalán nem	2 - kis mértékben	3 - jelentősen	4 rendkívül sokat
Egyetem: Alapképzés	0,00%	87,50%	12,50%	0,00%
Egyetem: Mesterképzés	0,00%	25,00%	50,00%	25,00%
Felső tagozat: 11. - 13. osztály	0,00%	33,33%	66,67%	0,00%
Felső tagozat: 7. - 10. osztály	0,00%	61,11%	38,89%	0,00%

21. ábra. A ChatGPT mennyire adott hasznos válaszokat a diákok szerint (táblázat)



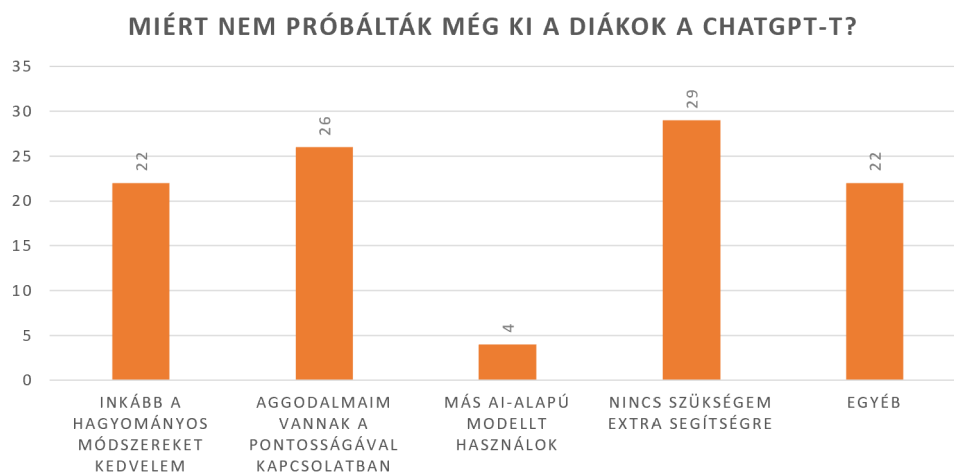
22. ábra. A ChatGPT mennyire adott hasznos válaszokat a diákok szerint (diagram)

A ChatGPT-t többféle célra szokták használni, leggyakrabban feladatok levezetésére és új fogalmak megértésére (lásd 23. ábra).



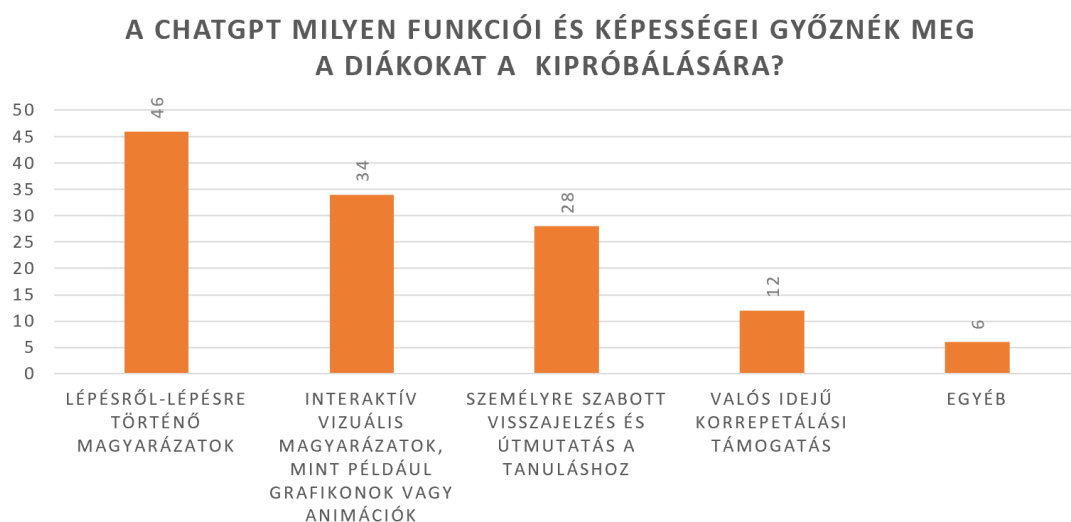
23. ábra. A ChatGPT használati céljai a diákok körében

A többsége azoknak, akik még nem próbálták ki, úgy gondolja, hogy nincs szükségük extra segítségre, vagy aggályaik vannak a pontosságával kapcsolatban; sokan az „Egyéb” kategóriában azt is megjegyezték, hogy még nem jutott eszükbe (lásd 24. ábra).



24. ábra. Okok, amiért a diákok még nem próbálták ki a ChatGPT-t

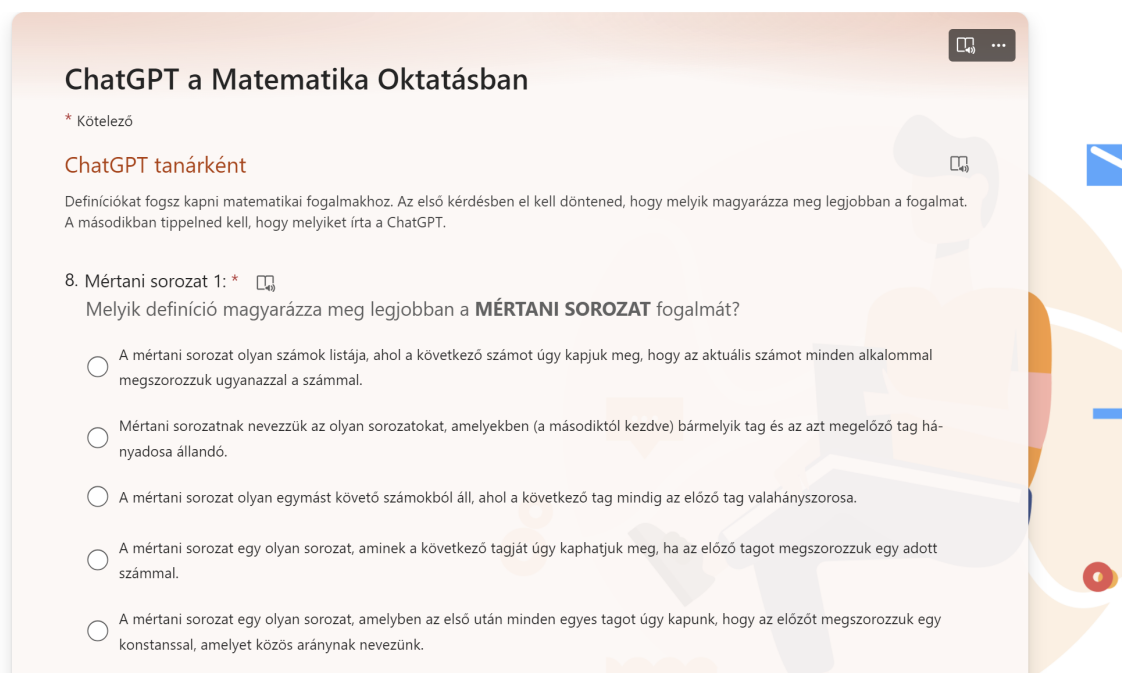
Amikor rákérdeztem, hogy milyen funkciókat találnának hasznosnak, a legtöbben a lépésről-lépésre történő magyarázatot jelölték meg (lásd 25. ábra), ennek érdekessége, hogy ezt már támogatják az újabb modellek, sőt tudományosan alátámasztott, hogy jobban is teljesítenek a modellek precíz érvelést igénylő feladatokban, ha elvárjuk tőlük, hogy lépésről lépésre oldják meg a feladatokat.



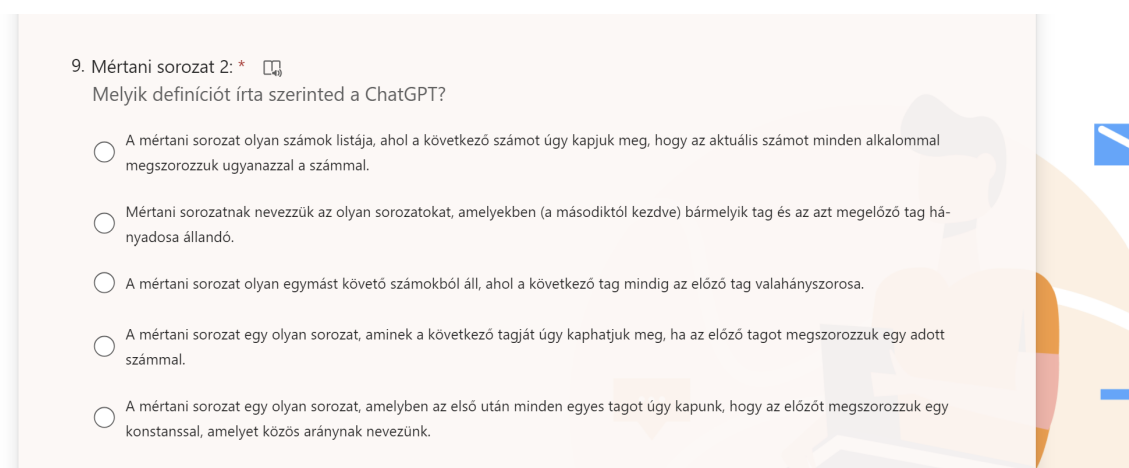
25. ábra. A diákok által kívánt új funkciók a ChatGPT-hez

4.2. Definíciók vizsgálata - ChatGPT mint tanár

Ebben a szakaszban Turing-tesztszerű kérdéseket tettem fel: négy fogalomhoz (mértani sorozat, logaritmus, határérték és derivált) öt-öt különböző definíciót adtam meg. Az öt definíció a következő forrásokból származott: középiskolai tanár, egyetemi matematika professzor, Wikipédia, valamint két ChatGPT-generált válasz, amelyek közül az egyik egy egyszerű definíció volt, a másik pedig egy professzori stílusú meghatározás. A válaszadókat arra kértem, hogy először jelöljék meg azt a definíciót, amelyik számukra a legérthetőbb, majd azt, amelyikről úgy gondolják, hogy a ChatGPT készítette (lásd 26. és 27. ábra).



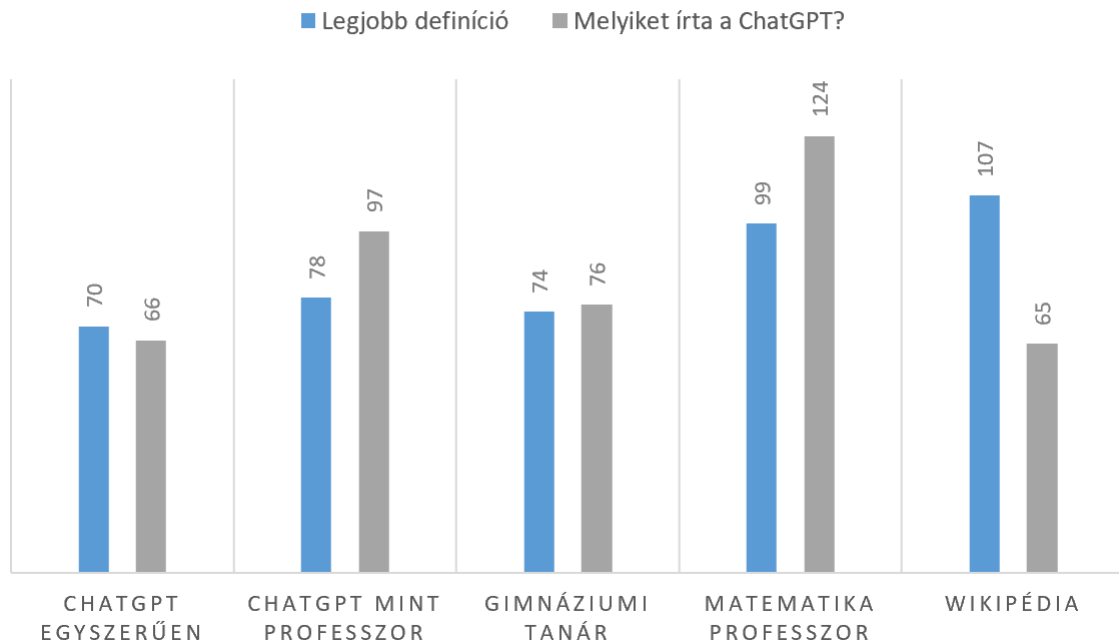
26. ábra. Minta a kérdőív második szakaszából (8. kérdés)



27. ábra. Minta a kérdőív második szakaszából (9. kérdés)

Az összesített eredmények alapján (lásd 28. ábra) a diákok a Wikipédia és a matematika professzor által írt definíciókat találták a legérthetőbbnek. Érdekes módon azonban a legtöbben a matematika professzor által írt definíciókat ismerték fel ChatGPT-eredetűként. Továbbá mátrixok segítségével megvizsgáltam, hogy a válaszadók az általuk legérthetőbbnek ítélt definíciót kinek tulajdonították. Az eredmények teljesen reprezentatívak voltak: akik például a professzor definícióját választották a legérthetőbbnek, arányosan jelölték a többi opciót a másik válaszadókkal egyetemben.

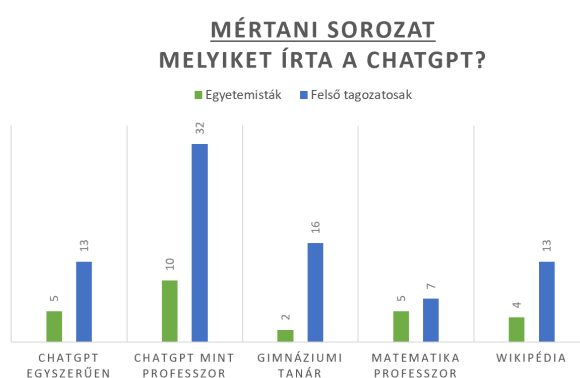
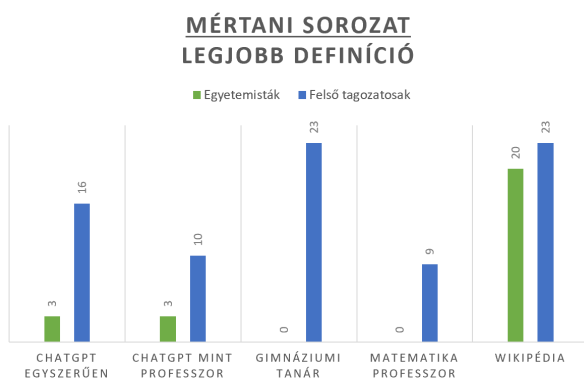
ÖSSZESÍTETT



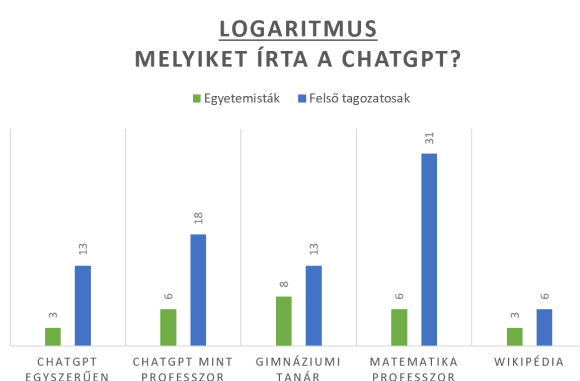
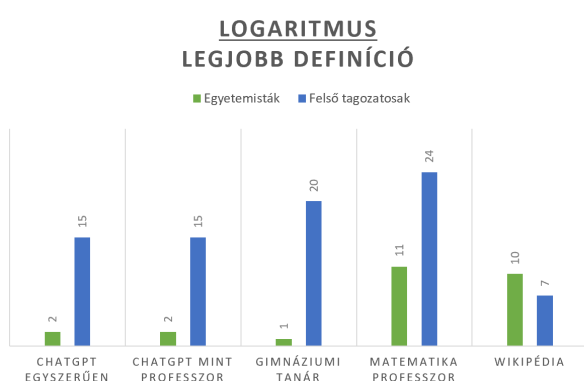
28. ábra. Összesített eredmények: Legjobb definíció és ChatGPT felismerése

Az ábrák jól mutatják, hogy a felső tagozatos diákok számára a ChatGPT által generált definíciók hasonlóan érthetőek voltak, mint bármely más forrásból származók, sőt a határérték fogalmának esetében a legtöbben a professzori stílusú ChatGPT-definícióra szavaztak. Az egyetemisták körében más eredmény született: számukra a matematika professzor és a Wikipédia által adott definíciók voltak a legérthetőbbek, és ritkábban választották a ChatGPT magyarázatait a legjobbnak.

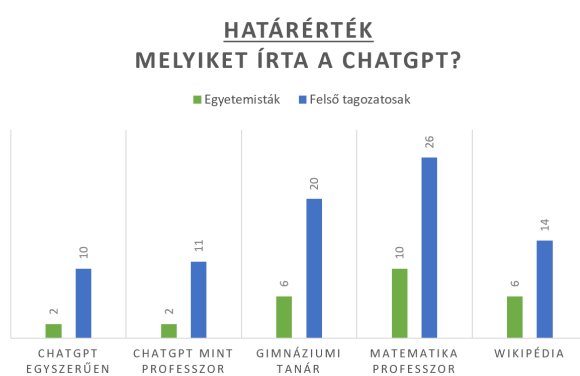
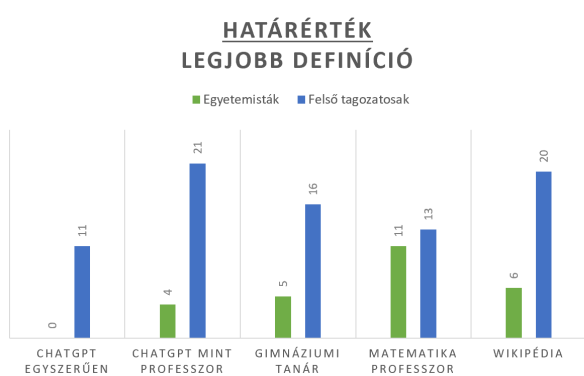
Amikor a diákoknak ki kellett választaniuk, hogy melyik definíciót készítette a ChatGPT, a hipotézisem beigazolódt: nem tudták konzisztensen megkülönböztetni a ChatGPT definícióit a többitől. A mértani sorozat esetében mind a középiskolások, mind az egyetemisták a professzori stílusú ChatGPT-magyarázatot jelölték meg, de a többi fogalomnál teljesen változó volt és többnyire a matematika professzor által írt definíciókat gondolták ChatGPT-eredetűnek (lásd 29-32. ábrák).



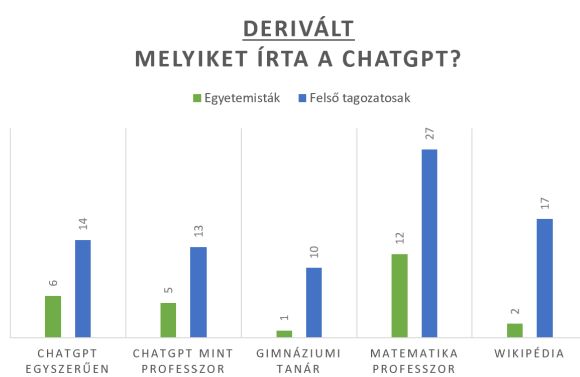
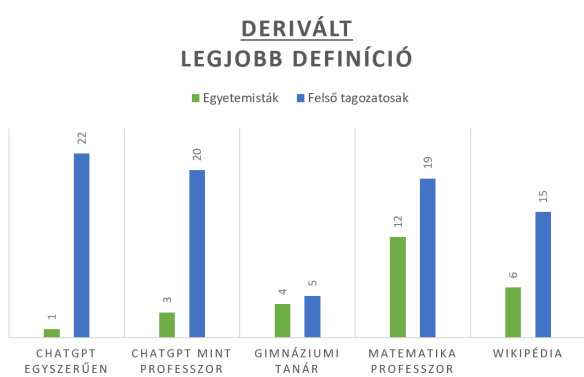
29. ábra. Mértani sorozat: Legjobb definíció és ChatGPT felismerés



30. ábra. Logaritmus: Legjobb definíció és ChatGPT felismerés



31. ábra. Határérték: Legjobb definíció és ChatGPT felismerés



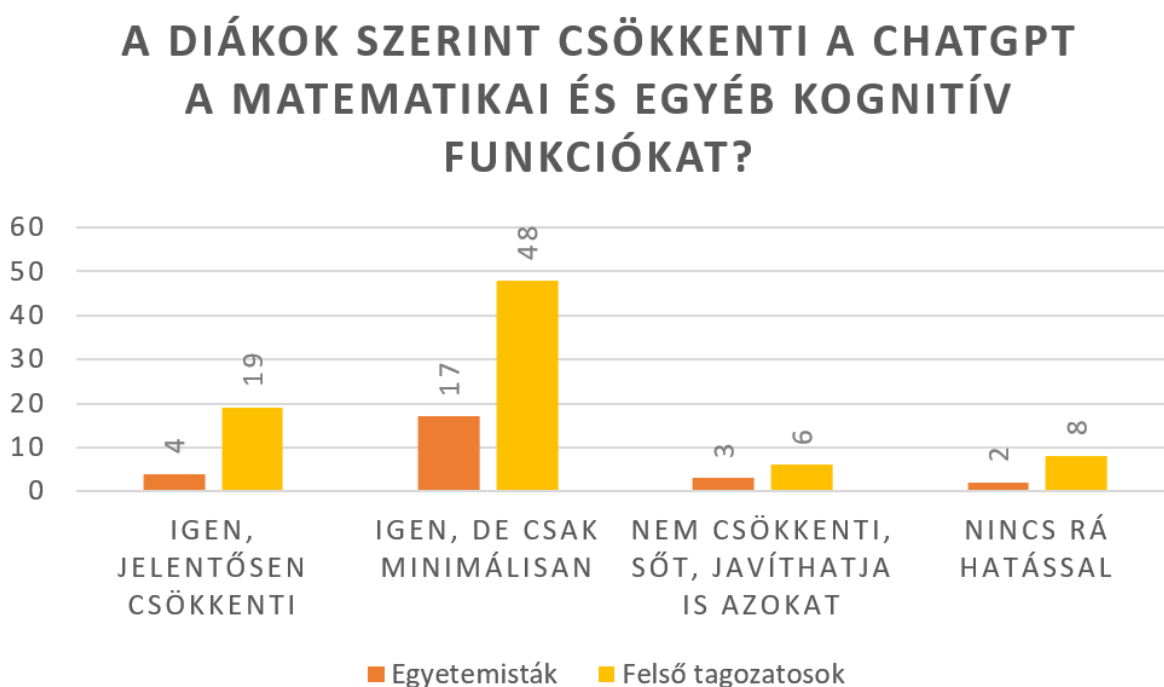
32. ábra. Derivált: Legjobb definíció és ChatGPT felismerés

4.3. Kitekintés - ChatGPT a jövőben

Ebben a szakaszban a ChatGPT jövőbeni hatását vizsgáltam a diákok véleménye alapján. A válaszok alapján mind a középiskolások, mind az egyetemi hallgatók többsége úgy véli, hogy a ChatGPT használata csökkentheti az ember matematikai és egyéb kognitív képességeit (lásd 33. és 34. ábrák). Ez azt jelenti, hogy a tudományos irodalomban kutatók által (például Heersmink munkásságában) jóslott hatás a közvéleményben is jelen van.

Évfolyam	Igen, jelentősen csökkenti	Igen, de csak minimálisan	Nem csökkenti, sőt, javíthatja is azokat	Nincs rá hatással
Egyetem: Alapképzés	15,79%	63,16%	15,79%	5,26%
Egyetem: Mesterképzés	14,29%	71,43%	0,00%	14,29%
Felső tagozat: 11. - 13. osztály	29,41%	47,06%	5,88%	17,65%
Felső tagozat: 7. - 10. osztály	21,88%	62,50%	7,81%	7,81%

33. ábra. A ChatGPT hatása a kognitív képességekre a diákok véleménye alapján (táblázat)

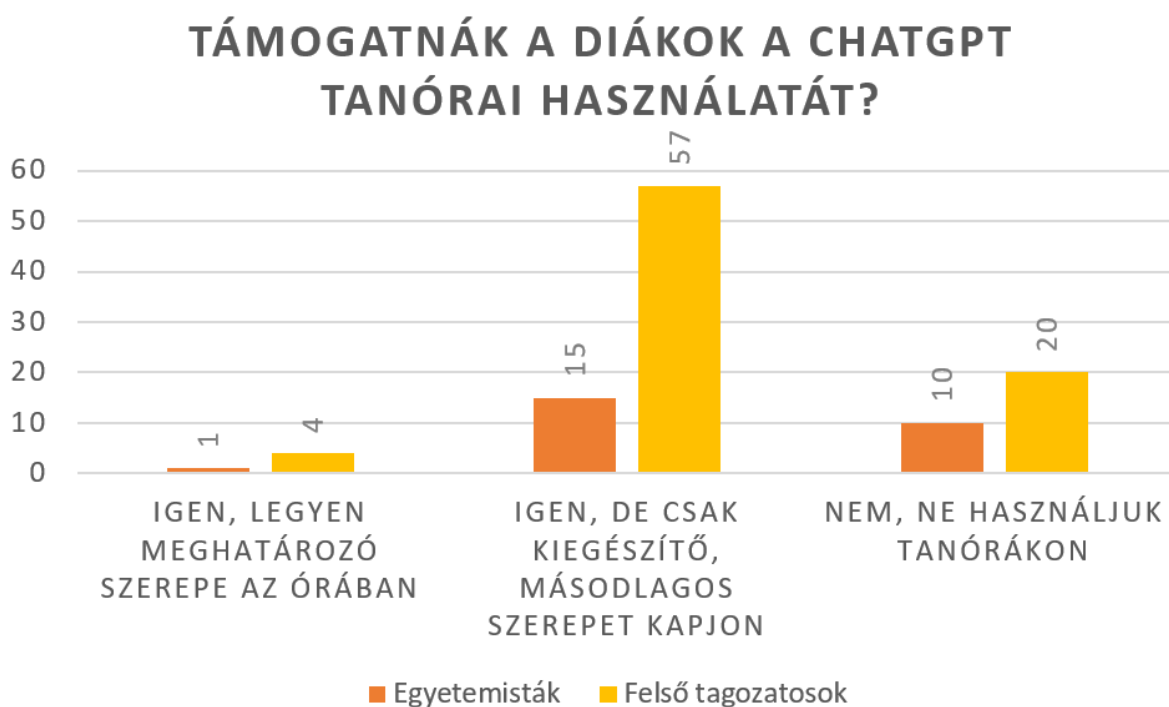


34. ábra. A ChatGPT hatása a kognitív képességekre a diákok véleménye alapján (diagram)

Ezzel együtt sok válaszadó, különösen a középiskolások, úgy gondolják, hogy a ChatGPT kiegészítő, másodlagos szerepet tölthetne be a tanórákon, ahol segítséget nyújthatna anélkül, hogy helyettesítené a tanárt. Az egyetemisták körében viszont megosztottabb vélemények születtek: több mint egyharmaduk úgy véli, hogy a ChatGPT-t egyáltalán nem kellene használni az oktatásban. Ez a hozzáállás arra utal, hogy míg sokan elismerik az AI hasznosságát, mégis fenntartásaik vannak a technológia oktatásba való beillesztésével kapcsolatban (lásd 35. és 36. ábrák).

Évfolyam	Igen, legyen meghatározó szerepe az órának	Igen, de csak kiegészítő, másodlagos szerepet kapjon	Nem, ne használjuk tanórákon
Egyetem: Alapképzés	0,00%	57,89%	42,11%
Egyetem: Mesterképzés	14,29%	57,14%	28,57%
Felső tagozat: 11. - 13. osztály	0,00%	70,59%	29,41%
Felső tagozat: 7. - 10. osztály	6,25%	70,31%	23,44%

35. ábra. A ChatGPT szerepe a tanórákon a diákok véleménye szerint (táblázat)

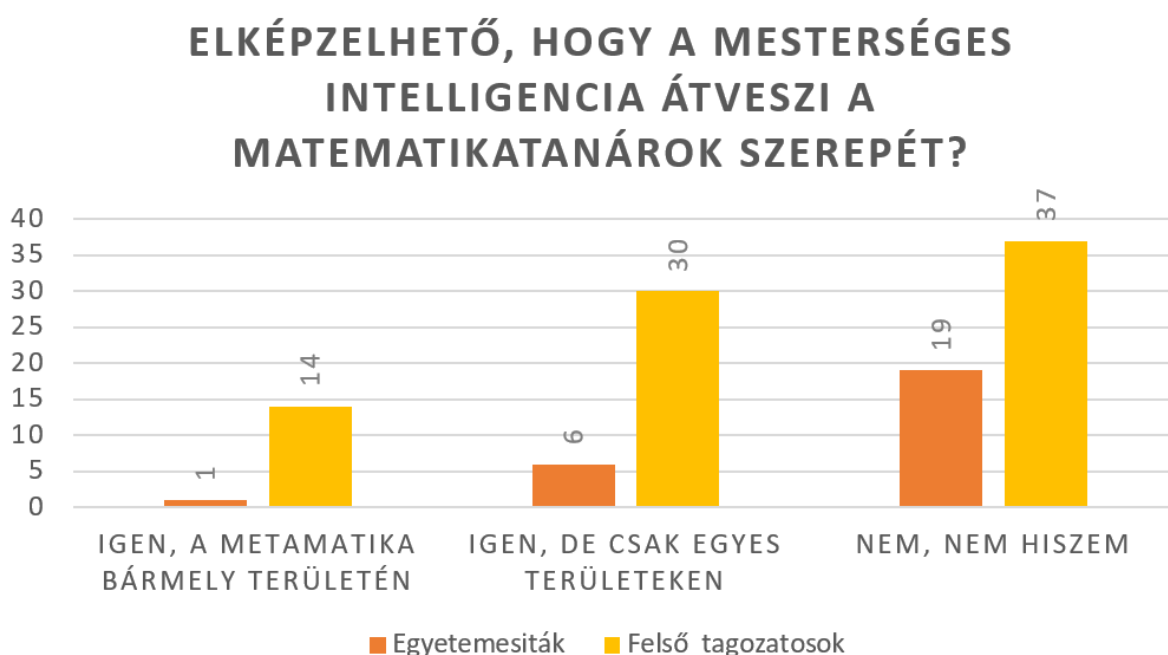


36. ábra. A ChatGPT szerepe a tanórákon a diákok véleménye szerint (diagram)

Az eredmények azt mutatják, hogy sokan kételkednek abban, hogy a mesterséges intelligencia teljes mértékben átveheti a matematikatanárok szerepét. Ez is azt mutatja, hogy az oktatásba való mélyebb integrációhoz még számos akadályt kell leküzdeni, és nem elegendő, ha az AI sikeresen old meg nehezebb feladatokat is – a diákok számára a tanári jelenlét továbbra is meghatározó és pótolhatatlan elem marad az oktatásban (lásd 37. és 38. ábrák).

Évfolyam	Igen, a matematika bármely területén	Igen, de csak egyes területeken	Nem, nem hiszem
Egyetem: Alapképzés	5,26%	21,05%	73,68%
Egyetem: Mesterképzés	0,00%	28,57%	71,43%
Felső tagozat: 11. - 13. osztály	0,00%	17,65%	82,35%
Felső tagozat: 7. - 10. osztály	21,88%	42,19%	35,94%

37. ábra. A mesterséges intelligencia képessége a tanárok helyettesítésére a diákok véleménye szerint (táblázat)



38. ábra. A mesterséges intelligencia képessége a tanárok helyettesítésére a diákok véleménye szerint (diagram)

A kérdőív végén a kitöltők szabadszöveges mezőben jelezhatték, ha van bármilyen további észrevételük, gondolatuk a mesterséges intelligencia és matematika, vagy matematika oktatás kapcsolatáról. Nagyon érdekes hasznos észrevételek érkeztek, melyek közül néhányat mindenképp érdemes megosztani. Az egyik gyakori észrevétel arról szólt, hogy hiába tudja a ChatGPT jól megoldani és elmagyarázni a feladatokat, a tanárnak nem csak ez a feladat. A tanárnak a diákok érdeklődését, motivációját is fenn kell tartania, sőt nevelő szerepe is van, mely egy chatboton keresztül semmiképpen sem oldható meg. Egy másik észrevétel arról szólt, hogy egyesek azért nem merik kipróbálni a ChatGPTt matematikatanuláshoz, mert attól félnek, hogy a chatbot nem értené meg, ha nem tudnák szakszavakkal elmagyarázni a feladatot. Ez egy jogos, bár alaptalan félelem, hiszen a ChatGPTnek pont ez az egyik nagyon jó tulajdonsága, hogy vég-

telen türelemmel rendelkezik, és saját szavainkkal körülírva is mindig megpróbálja megérteni a problémát, és segíteni annak megoldását. Végül, de nem utolsó sorban egy jogos aggály a ChatGPT alkalmazhatóságával kapcsolatban, hogy igazán akkor tudja segíti nagyon jól diákok munkáját, ha nem nulláról indulnak az adott témakörben, hanem rendelkeznek már egy alap tudással. Ugyanis előfordulhat és elő is fog fordulni, hogy a chatbotok hibáznak, és abban az esetben, ha a területhez abszolút nem értünk, akkor könnyedén becsaphatnak minket, tehát rendelkezniünk kell a megfelelő használatukhoz egy kellő alaptudással, mellyel fel tudjuk ismerni, ha logikai buktatók vagy egyéb hibák vannak a chatbot válaszaiban.

5. Zárógondolatok

Az irodalomkutatásomban néhány úttörő cikk alapján áttekintettem a nagy nyelvi modellek és a matematika kapcsolatának jelenlegi helyzetét. Az AlphaGeometry modell, amelyet Terence Tao is hatalmas potenciállal rendelkezőként említ, megmutatta, hogy a mesterséges intelligencia alapú modellek a bizonyítást segítő rendszerekkel együttműködve akár a világ legjobb matematika olimpiikonjaival is versenyre kelhetnek. A Qwen2 csapata pedig egy olyan nyílt forráskódú modellt hozott létre, amely matematikában és kódolásban felülmúlja a korábban legjobbnak tartott Gemini-Math-1.5-Pro és GPT-4o modelleket.

Mivel ezek a modellek még nem elérhetők mindennapi használatra – mert nem publikusak, vagy csak több tízmilliós számítógépeken tudnánk futtatni őket –, elemzésemben a két legismertebb és szintén kiváló teljesítményű modell, a GPT-4o és a GPT-o1 teljesítményét vizsgáltam. Személyes tapasztalataim és egy kiértékelés alapján igazoltam, hogy ezek az új modellek a matematikában nyelvtől függetlenül szinte azonos szinten teljesítenek. Kiértékeltem a modellek teljesítményét az elmúlt öt év közép- és emelt szintű matematika érettségi feladatain, valamint megvizsgáltam a KöMaL feladatokon is. A GPT-4o modell ma már közép- és emelt szinten is jelesre teljesítene, míg a GPT-o1 modell – amely Terence Tao szerint egy átlagos mesterszakos hallgató szintjén áll – emelt szinten 90 százalék feletti eredményt érne el. A GPT-o1 modell teljesítményét a KöMaL feladatain is elemeztem: C-jelű feladatoknál kielégítően teljesít, de B- és A-jelű feladatoknál már kevésbé hatékony. Ez jelentős előrelépés az elmúlt évekhez képest, hiszen a ChatGPT korábbi modellje, amely ma már nem elérhető, még alapvető szorzásokat is gyakran elhibázott. Az általános tendencia azonban egyértelmű: egyre erősebb modellek készülnek, és az OpenAI ígérete szerint a fejlesztés alatt álló GPT-o1 modellt hamarosan PhD szintű tudással ruházzák fel.

Ezzel a kutatással megtettük az első kritikus lépést annak érdekében, hogy az AI rendszerek valóban integrálhatók legyenek az oktatásba. Ennek feltétele, hogy alaposan megismerjük a mesterséges intelligencia lehetőségeit és korlátait, hiszen anélkül nem tudnánk megfelelően használni őket. Az eredmények alapján kiderült, hogy ezek a modellek – különösen az o1 modell – nemcsak az egyszerű, hanem a bonyolultabb feladatok megoldásában is kiváló teljesítményt nyújtanak. Ezért, ha egy diáknak segítségre lenne szüksége, ezek az AI modellek

versenyképes támogatást tudnak biztosítani, ami hasznos eszközzé teheti őket az oktatásban.

A kérdőív válaszai azt is megmutatták, hogy a modell jól boldogul az absztrakt fogalmak definiálásával. Az általa adott magyarázatokat sokan a legérthetőbbnek találták, és ezek a definíciók nem tűntek ki úgy, mintha „gépi származékok” lennének, ami különösen pozitív visszajelzés. Emellett az is kiderült, hogy a középiskolás diákok alapvetően nyitottak arra, hogy a ChatGPT szerepet kapjon a matematikaórákon. Ez a tanárok számára egy új lehetőséget kínálhat arra, hogy a mesterséges intelligenciát támogató eszközként alkalmazzák az órák során, segítve ezzel a diákok megértését és fejlődését a tantárgyban.

Ezt a kutatást a jövőben több irányban is érdemes lenne folytatni, hiszen számos további kérdés és lehetőség rejlik benne. Érdekes lenne például összehasonlítani a ChatGPT teljesítményét más mesterséges intelligencia modellekkel szemben, mint például a Claude.ai vagy Gemini vagy a nyíltforráskódú Qwen2 modellek, hogy megtudjuk, melyik modell oldja meg sikeresebben a különböző feladatokat. Sőt, amennyiben még nagyobb erőforrás állna rendelkezésünkre megpróbálhatnánk az AlphaGeometryhoz hasonló rendszert építeni mondjuk KöMal feladatokhoz, hátha egy KöMal matematikaversenyekre specializálódott mesterséges intelligencia már jó teljesítményt érne el a B- és A-jelű feladatokban is. Emellett a kérdőív is bővíthető lenne: kialakíthatnánk egy új szakaszt, ahol nemcsak definíciókat, hanem konkrét feladatok részletes levezetéseit is bemutatnánk a kitöltőknek, majd megkérnénk őket, hogy értékeljék azokat. Ezzel pontosabb képet kapnánk arról, hogy a diákok és felhasználók mennyire találják érthetőnek és hasznosnak a modellek által adott részletes, lépésről-lépésre történő magyarázatait, ami értékes visszajelzést nyújtana a további fejlesztésekhez.

Hivatkozások

- Dinh, T. A., Mullov, C., Bärman, L., & et al. (2024). Sciex: Benchmarking large language models on scientific exams with human expert grading and automatic grading. (<https://arxiv.org/abs/2406.10421>)
- Gemini Team Google. (2024). Gemini: A family of highly capable multimodal models. (<https://arxiv.org/abs/2312.11805>)
- Google cloud: Dialogflow. (2023). (<https://cloud.google.com/dialogflow>)
- Heersmink, R. (2024). Use of large language models might affect our cognitive skills. *Nature Human Behaviour*, 8, 805–806.
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8).
- Imani, S., Du, L., & Shrivastava, H. (2023). Mathprompter: Mathematical reasoning using large language models. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 5, 37–42.

- Lo, C. K. (2023). What is the impact of chatgpt on education? a rapid review of the literature. *Education Sciences*, 13(4).
- OpenAI. (2024). Gpt-4 technical report.
(<https://arxiv.org/abs/2303.08774>)
- Qwen Team. (2024). Qwen2 technical report.
(<https://arxiv.org/abs/2407.10671>)
- Tao, T. (2024a, June 8). *Ai will become mathematicians' co-pilot*. Retrieved from <https://www.scientificamerican.com/article/ai-will-become-mathematicians-co-pilot/>
- Tao, T. (2024b, October 4). *Interview in the atlantic*. Retrieved from <https://www.theatlantic.com/technology/archive/2024/10/terence-cao-ai-interview/680153/>
- Topsakal, O., & Topsakal, E. (2022). Framework for a foreign language teaching software for children utilizing ar, voicebots and chatgpt (large language models). *The Journal of Cognitive Systems*, 7(2).
- Trinh, T. H., Wu, Y., Le, Q. V., He, H., & Luong, T. (2024). Solving olympiad geometry without human demonstrations. *Nature*, 625, 476–482.
- Vaswani, A., Shazeer, N., Parmar, N., & et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 6000 - 6010.