ESZTERHÁZY KÁROLY KATOLIKUS EGYETEM
INFORMATIKA KAR • DIGITÁLIS TECHNOLÓGIA INTÉZET
AGRIA MÉDIA KONFERENCIA 2025

https://doi.org/10.17048/AM.2025.209

https://www.videotorium.hu/hu/recording/56486?playlist=5954

Ujhelyi Gábor
Számonkérések automatikus értékelési lehetősége nagy nyelvi modellek felhasználásával

Ujhelyi Gábor

Eszterházy Károly Katolikus Egyetem

ug@mensa.hu

Absztrakt: A mesterséges intelligencia – köztük a nagy nyelvi modellek (LLM-ek) – gyors fejlődése érdemben alakítja az oktatási értékelés területét: a részfolyamatok automatizálásának ígérete mellett új kockázatokat és módszertani kérdéseket is felvet. Kutatásomban az instruktív, lépésről lépésre megfogalmazott Microsoft Word-szövegszerkesztési feladatok automatikus értékelhetőségét vizsgáltam felhőszolgáltatásként elérhető LLM-ekkel. Hallgatóim ZH-ként beadott megoldásait először manuálisan, majd több iterációban különböző promptstratégiákkal LLM-ek segítségével értékeltem, és az eredményeket egymással, illetve a manuális pontozással hasonlítottam össze. A vizsgálat alapján az LLM-ek a viszonylag egyszerűbb formázási követelményeket többnyire megbízhatóan észlelik, ám a komplexebb, funkcionális elemek (pl. tartalomjegyzék, ábrajegyzék, irodalomjegyzék) esetén gyakoriak a hamis pozitív és hamis negatív döntések. A modellek közül a ChatGPT és a Claude eredményei közelítettek leginkább a manuális értékeléshez; a Gemini, a Perplexity és a DeepSeek sokszor túlzottan optimistán pontozott, míg a Manus teljesítménye maradt el a leginkább. Következtetésem szerint az LLM-ek önálló, zárt láncú értékelésre még nem alkalmasak, de jól körülhatárolt részfeladatokra, a tanári döntést támogató eszközként már most hasznosíthatók. Eredményeim megerősítik, hogy a promptolás strukturálása („lépésenkénti” gondolatvezetés) javíthatja a konzisztenciát, ugyanakkor a nyelvi megfogalmazásra – az instrukciók nyelvére és precizitására – érzékeny a teljesítmény. A továbblépéshez több-lépéses feldolgozási csővezetékek, integrációs megoldások és további modellvariánsok vizsgálatát tartom ígéretesnek.

A Kulturális és Innovációs Minisztérium EKÖP-24 kódszámú Egyetemi Kutatói Ösztöndíj Programjának a Nemzeti Kutatási, Fejlesztési és Innovációs Alapból finanszírozott szakmai támogatásával készült.

Kulcsszavak: nagy nyelvi modell, automatikus értékelés, szövegszerkesztés, oktatástechnológia, felsőoktatás

Automatic evaluation of assessments using large language models

Abstract: The rapid development of artificial intelligence – including large language models (LLMs) – is significantly changing the field of educational assessment: in addition to the promise of automating sub-processes, it also raises new risks and methodological issues. In my research, I examined the automatic assessment of instructive, step-by-step Microsoft Word word processing tasks using LLMs available as a cloud service. I first assessed my students' solutions submitted as ZH manually, then in several iterations using different prompting strategies using LLMs, and compared the results with each other and with manual scoring. Based on the study, LLMs mostly reliably detect relatively simple formatting requirements, but false positive and false negative decisions are common in the case of more complex, functional elements (e.g. table of contents, list of figures, bibliography). Of the models, the results of ChatGPT and Claude came closest to manual assessment; Gemini, Perplexity and DeepSeek often scored too optimistically, while Manus' performance was the most lacking. I

conclude that LLMs are not yet suitable for independent, closed-loop assessment, but they can already be used for well-defined subtasks, as a tool to support teacher decisions. My results confirm that structuring prompting ("step-by-step" thought guidance) can improve consistency, but performance is sensitive to linguistic formulation – the language and precision of the instructions. I consider the examination of multi-step processing pipelines, integration solutions and further model variants to be promising for further progress.

Keywords: large language model, automatic assessment, text editing, educational technology, higher education

1 Bevezetés

A generatív mesterséges intelligencia (MI) megjelenése és gyors innovációs ciklusai új korszakot nyitottak az oktatásban: egyszerre nyújtanak példátlan lehetőségeket a tanulási folyamatok személyre szabására és az értékelés támogatására, ugyanakkor új, megoldandó módszertani és etikai kérdéseket is felvetnek (Santos & Junior, 2024; Denecke, Glauser, & Reichenpfader, 2023). A nemzetközi szakirodalomban az automatikus itemgenerálás és feleletválasztós tesztek (MCT) minőségbiztosítása, illetve a szabad szövegek automatikus pontozása (Automated Essay Scoring, AES) már évtizedek óta kutatott területek (Mead & Zhou, 2023; Shermis, 2014; Bai et al., 2022). Ezzel szemben a dokumentumszerkesztési feladatok (például a Microsoft Word funkcionális és tipográfiai követelményeinek ellenőrzése) kisebb figyelmet kaptak, pedig mind a felsőoktatásban, mind a munkaerőpiacon alapkompetenciának számítanak (Kurniawan et al., 2024).

A most bemutatott kutatás Word-alapú, részfeladatokra bontott ZH-k automatikus értékelhetőségét vizsgálta felhőben elérhető LLM-ek segítségével. A módszertani keretben a manuális pontozást tekintettem referenciának, majd különböző modellekkel és promptstratégiákkal generált automatikus értékeléseket hasonlítottam ehhez. Célom nem egy végleges automatizált bírálati rendszer létrehozása volt, hanem a lehetőségek és korlátok feltárása, különös tekintettel a részfeladatok típusára és az instrukciók megfogalmazására. A vizsgálat tárgyát képező kurzusok az Eszterházy Károly Katolikus Egyetem Információs és Kommunikációs Technológiák című tantárgyához kapcsolódtak, a hallgatói megoldásokat valós ZH-ként gyűjtöttem be és értékeltem.

2 A kutatás előzményei

Az MI-alapú értékelés történeti ívét tekintve a korai rendszerek jellemzően szabályalapúak vagy hagyományos gépi tanulási módszerekre épültek, az AES hagyománya jól dokumentált, és összehasonlító vizsgálat járult hozzá a terület fejlődéséhez (Shermis, 2014). Az LLM-ek megjelenése ugyanakkor új képességeket hozott, a természetes nyelvű instrukciók megértését, magyarázatok generálását, valamint a heterogén bemenetek kezelését, amelyeket pedagógiai és mérés-értékelési kontextusokban is vizsgálnak (Bai et al., 2022; Santos & Junior, 2024).

A vizsga- és tesztfejlesztés oldalán a generatív modellek ígéretesek az itemek automatikus létrehozásában és kalibrálásában (Mead & Zhou, 2023), míg a tanulói nyílt végű válaszok értékelésében vegyes eredmények láthatók, az LLM-ek gyakran képesek releváns szempontok azonosítására, de hajlamosak a túlzott magabiztosságra és a konzisztenciahiányra, különösen strukturált, szabálykövető feladatok esetén (Bai et al., 2022; Kurniawan et al., 2024). A felsőoktatási gyakorlatban a lehetőségek mellett a kockázatokat (például a torzítások, a megbízhatóság és az átláthatóság kérdését) rendszeresen kiemeli az empirikus és áttekintő irodalom (Denecke et al., 2023; Santos & Junior, 2024).

Ezen történeti és elméleti háttérre támaszkodva a kutatásom az LLM-ek dokumentumszerkesztési kompetenciák értékelésében betöltött szerepét vizsgálja, azaz, hogy milyen feltételek mellett, milyen részfeladatoknál és milyen promptstratégiákkal közelíthető meg a manuális pontozáshoz hasonló döntési szint.

3 A kutatás aktualitása

Az oktatás digitális transzformációja, az online tanulási környezetek és a hallgatói létszámok növekedése az értékelés hatékonyságának kérdéseit a figyelem középpontjába emelte. Az LLM-ek elterjedése (az informatikai infrastruktúra bővülésével és a felhasználói szokások átalakulásával együtt) sürgetővé teszi, hogy a felsőoktatás differenciált, adaptív és gyors visszajelzést nyújtson, ugyanakkor a megbízhatóságot és a méltányosságot fenntartsa. A vizsgálatom közvetlen motivációja ezért kettős, egyrészt kvantifikálni szerettem volna, hogy a mai, felhőben elérhető LLM-ek miként teljesítenek konkrét, Wordön belül megvalósítandó részfeladatok ellenőrzésében, másrészt feltárni, hogyan befolyásolja a promptolás szerkezete és pontossága az eredményeket.

A kurzusaimban a szövegszerkesztési készségek nem pusztán formázási megjelenést jelentenek, hanem a dokumentumfunkciók (például stílusalapú címsorok, automatikus tartalomjegyzék, idézetkezelés és irodalomjegyzék) szakszerű használatát. A funkcionális elemek validálása nagyobb kihívást jelent a modelleknek, mint az egyszerű attribútumok (betűtípus, igazítás), ezért a vizsgálat aktualitását a tanulói és tanári munkaterhelés csökkentésének, valamint a visszajelzési ciklus lerövidítésének igénye is erősen támogatja (Santos & Junior, 2024; Bai et al., 2022).

4 A kutatás megvalósítása

4.1 Mintázat és feladatarchitektúra

A vizsgálat alapját az Eszterházy Károly Katolikus Egyetem IKT tárgyaihoz kapcsolódó, két szemeszteren át futó kurzusok adták. Összesen 119 hallgató vett részt a felmérésben, akik közül hiányzások és technikai okok miatt 107 dolgozat került ténylegesen értékelésre. A hallgatók ZH-ként beadott Word dokumentumai 26 részfeladatot tartalmaztak, melyek a formázási és funkcionális kompetenciák széles körét fedték le, úgy, mint címszintek, felsorolások, betűattribútumok, igazítás, idézetek, valamint automatikus tartalomjegyzék-, ábrajegyzék- és irodalomjegyzék-készítés. A feladatsort kísérő mintamegoldás PDF formájában is rendelkezésre állt, a megoldásra 100 perc állt rendelkezésre (a feladatsor ideális esetben 20 perc alatt is megoldható).

4.2 Értékelési eljárás és összehasonlítás

Az értékelés két lépcsőben történt. Először részfeladatonként manuálisan pontoztam a beadott dokumentumokat, referenciaként szolgálva az automatikus értékelésekhez. Ezt követően felhőszolgáltatásként elérhető LLM-ekkel végeztem a kiértékelést, a következő modellekkel: ChatGPT 4o, DeepSeek R1, Claude, Gemini, Perplexity, valamint Manus. A dolgozatokat egyenként töltöttem fel és értékeltettem a modellekkel (kivéve Manus), majd a modellek eredményeit egymással és a manuális pontozással vetettem össze.

A promptolási stratégiák több iterációban finomodtak. Kipróbáltam a tömör, „csak pontozás” jellegű instrukciót, visszajelzés-alapú újraértékeltetést (az előző körben detektált hibák explicit közlésével), a pontozás szabályosságára vonatkozó hangsúlyokat, a lépésekre bontott, sorrendet és ellenőrző listát követő megközelítést (lépésenkénti gondolatvezetés), végül az eredmények táblázatos összesítését kérő változatokat.

4.3 Mérési megfigyelések és gyakorlati korlátok

A gyakorlatban az ingyenes hozzáférési limitek szűknek bizonyultak, a modellek felhasználói felület alapú kvótái a minta mennyiség egyidejű feldolgozását megnehezítették. A legkövetkezetesebb teljesítményt a ChatGPT mutatta, amelyhez a Claude eredményei hasonlítottak, ugyanakkor a Claude néhol konzervatívabb („negatívabb”) pontozási hajlamot mutatott. A Gemini, a Perplexity és a DeepSeek több esetben túlzottan optimista képet adott a dolgozatokról, azaz az instrukcióknak nem megfelelő elemeket a valósnál nagyobb arányban minősítette helyesnek. A magáról komplex feladatok megoldását ígérő Manus teljesítménye jelentősen elmaradt mind az előzetes várkozásoktól, mind a többi modell teljesítményétől. A korrelációs mintázatok arra utaltak, hogy az

értékelhető modellek egymással és a manuális pontozással is érzékelhető együttjárást mutattak, bár a torzítások típusa és mértéke modellenként eltért.

A promptok hatásai egyértelműek voltak. A tömör instrukciók gyakran túlzottan derűlátó, „jóindulatú" értékeléshez vezettek, amikor azonban visszakérdezéses pontosítást kértem, előfordult, hogy a modell „átbillent" és indokolatlanul sok hibát vélt felismerni. A teljesen hiányzó elemeket a modellek többnyire jól azonosították, de a funkcionális, dokumentum-szintű elemek (például a stílus-alapú címsorok megléte, az automatikus tartalomjegyzék korrekt konfigurálása vagy az irodalomjegyzék beállításainak megfeleltetése) gyakran okoztak gondot. Mindkét domináns modell (ChatGPT, Claude) relatíve következetesebb teljesítményt nyújtott a többiekhez képest.

4.4 Megbízhatóság, nyelvi tényezők és metodikai tanulságok

Az összevetés alapján az LLM-ek önálló, oktatói beavatkozást nem igénylő értékelőrendszerként a kutatás végrehajtásakor elérhető fejlettségi szintjükön nem tekinthetők elégségesnek. A modellválasztás és feladattípus erősen befolyásolja a pontosságot, a „chain-of-thoughts" jellegű, lépésenkénti promptok előnyt mutattak, különösen, ha az ellenőrzési lépések explicit listaként szerepeltek. Összességében fejlettebb nyelvi modellre épülő, és/vagy komplexebb (több lépésből, ellenőrző pontokból és esetleg külső eszköz-integrációból álló) megoldás szükséges a megbízható automatizált értékeléshez.

5 Eredmények és értelmezés

5.1 Modell összehasonlítás, összkép

A modellek viselkedésében három karakteres mintázat rajzolódott ki. Először is, a ChatGPT és a Claude a részfeladatok többségében közelebb került a manuális pontozáshoz, ez a stabilitás a promptstruktúrára is kevésbé volt érzékeny, bár a lépésenkénti megközelítés itt is javító hatású volt. Másodszor, a Gemini, a Perplexity és a DeepSeek rendszeresen alulbecsülte a hibákat, például helyesnek minősítették azokat a dokumentumokat is, amelyekben formailag felülírt címsorok voltak stílusok helyett, vagy a tartalomjegyzék nem automatikus eszközzel készült. Harmadszor, a Manus esetén az alacsonyabb teljesítmény mellett a következetlenség is hangsúlyosabban jelent meg.

Ezek a mintázatok összhangban állnak azzal a szakirodalmi képpel, amely szerint az LLM-ek (noha képesek komplex instrukciók értelmezésére) a strukturált, szabályvezérelt validálási feladatokban (pl. formális dokumentumfunkciók ellenőrzése) hajlamosak heurisztikákra és felszíni jegyekre támaszkodni, ami a konzisztenciát rontja (Bai et al., 2022; Kurniawan et al., 2024).

5.2 Feladattípusok és hibaprofilok

A legegyszerűbb attribútumok (betűméret, félkövér/kurzív jelenléte, bekezdés-igazítás) esetén a helyes felismerési arányok magasak voltak. Ezzel szemben a dokumentumszintű, funkcionális követelményeknél (pl. a címszintek stílusokkal való kontrollált használata, a tartalomjegyzék generálása és frissíthetősége, idézetek és irodalomjegyzék konzisztenciája) gyakoriak voltak a téves értékelések. Tipikus hamis pozitív eset, amikor a modell pusztán vizuális hasonlóság alapján helyesnek minősítette a manuálisan formázott címsorokat, miközben hiányzott a stílusbeállítás.

A visszajelzéses újraértékeltetés kettős élűnek bizonyult, a részletes hibajelzés javította ugyan a modellek figyelmét, de az „átbillenés" jelensége (a túlzott hibakeresés) több esetben torzított. Ez a dinamika egybecseng azzal a tapasztalattal, hogy az LLM-eknél a hibatípusokra való fókuszálás a döntési küszöböket is elmozdítja, miközben nő a magyarázó szövegek hossza és bizonytalansága.

5.3 Promptstratégia és módszertani hatások

A strukturált, lépésről lépésre vezetett promptok (ellenőrző-listával és kötelező megállókkal) egyértelműen javították az értékelés következetességét. A pontosság ugyanakkor nem nőtt minden feladattípusnál egyformán, a vizuális attribútumoknál kisebb, a funkcionális elemeknél nagyobb nyereség volt mérhető.

Összességében az LLM-ek automatikus értékelési alkalmazása a vizsgált feladatkörben jelenleg olyan hibrid megoldások irányába mutat, amelyekben a modell(ek) előszűrnek, javaslatot tesznek, vagy magyarázatot generálnak, miközben a végső döntés oktatói kontroll alatt marad. Ez egybecseng a felsőoktatási gyakorlatban megfogalmazott ajánlásokkal, amelyek a tanári szerepet támogatói irányba viszik, ahelyett, hogy teljesen kiváltanák (Santos & Junior, 2024; Denecke et al., 2023).

6 Következtetések és további kutatási lehetőségek

A kutatási eredmények alapján három fő következtetés vonható le. (1) Az LLM-ek jelenleg nem alkalmasak önálló értékelésre Word-feladatok esetén a megbízhatósági és konzisztencia tapasztalatok alapján. (2) A feladat-specifikus, funkcionális elemek ellenőrzése összetett probléma, amely az LLM-ek felszíni mintafelismerési hajlamai miatt különösen sérülékeny. (3) Jól körülhatárolt, egyszerűbb részfeladatokra (például attribútumok ellenőrzésére vagy hiányzó elemek detektálására) a modellek már most is érdemben használhatók a tanári munka támogatására, különösen strukturált promptokkal.

A továbblépés, folytatás illetve kapcsolódó kutatási területek, amelyeket a kutatásom is körvonalaz, a következő irányokat tartalmazza: (1) további modellvariánsok és frissítések vizsgálata; (2) több-lépéses feldolgozási pipeline kialakítása (pl. dokumentum-metaadatok és stílusobjektumok explicit kiolvasása, szabályalapú ellenőrző modulokkal kombinálva); (3) táblázatkezelési feladatok hasonló elemzése; (4) értékelési platformokba ágyazott integrációk és felhasználói felületek; (5) célfeladatokra hangolt, akár saját modell(ek) finomhangolása; (6) ügynökhálózatok (agent networks) vizsgálata, amelyekben a feladatot több specializált komponens oldja meg együttműködésben.

Ezen javaslatok illeszkednek a nemzetközi trendekhez, a szakirodalom is a hibrid, többkomponensű megoldásokat tartja ígéretesnek, amelyek az LLM-ek rugalmas nyelvi megértését a szabályalapú és formális ellenőrzések determinisztikusságával ötvözik (Bai et al., 2022; Kurniawan et al., 2024).

7 Összegzés

Kutatásomban egy felsőoktatási környezetben, valós ZH-kra támaszkodva mértem fel, hogy a felhőben elérhető LLM-ek milyen mértékben képesek a Microsoft Word szövegszerkesztési feladatok automatikus értékelésére. A manuális értékelést referenciaként használva azt találtam, hogy a modellek közül a ChatGPT és a Claude teljesítménye közelít a leginkább az emberi bírálathoz, míg a Gemini, a Perplexity és a DeepSeek inkább túlértékel, a Manus pedig jelentős lemaradást mutat. A promptstratégia lényeges, a lépésenkénti, ellenőrző listát követő instrukciók láthatóan javítják a konzisztenciát, de a funkcionális dokumentumelemek helyes megítélésében így is gyakoriak a téves döntések. Konklúzióm szerint ilyen jellegű feladatok területén az LLM-ek jelenleg csupán tanári döntést támogató eszközként kínálnak hozzáadott értéket (például előszűrésre, hiányok detektálására vagy magyarázó visszajelzésre), míg a nagyobb jelentőségű, illetve végleges értékelést egyelőre fontos emberi kontroll alatt tartani.

Irodalomjegyzék

Bai, J. Y. H., Zawacki-Richter, O., Bozkurt, A., Lee, K., Fanguy, M., Sari, B. C., & Marín, V. I. (2022). Automated Essay Scoring (AES) Systems: Opportunities and Challenges for Open and Distance Education. *Tenth Pan-Commonwealth Forum on Open Learning (PCF10)*.

Denecke, K., Glauser, R., & Reichenpfader, D. (2023). Assessing the potential and risks of AI-based tools in higher education: Results from an eSurvey and SWOT analysis. *Trends in Higher Education, 2*(4), 667–688. https://doi.org/10.3390/trends2030040

Kurniawan, W., Riantoni, C., Lestari, N., & Ropawandi, D. (2024). A hybrid automatic scoring system: Artificial intelligence–based evaluation of physics concept comprehension essay test. *International Journal of Information and Education Technology, 14*(6), 786–882.

Mead, A., & Zhou, C. (2023). Evaluating the quality of AI-generated items for a certification exam. *Journal of Applied Testing Technology, 24*.

Santos, S. C. dos, & Junior, G. A. (2024). Opportunities and challenges of AI to support student assessment in computing education: A systematic literature review. In *Proceedings of the 16th International Conference on Computer Supported Education (CSEDU 2024)* (Vol. 2, pp. 15–26). https://doi.org/10.5220/00125525

Shermis, M. D. (2014). State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration. *Assessing Writing, 20*, 53–76. https://doi.org/10.1016/j.asw.2014.03.002

Tóthné P. L., Lengyelné M. T., Kis-Tóth L. (2014): Statisztikai programrendszerek. Eger, Líceum Kiadó. ISBN: 9786155250842