Mi történik, ha a ChatGPT-t bízod meg hatástanulmányok értékelésével?[1]

Sasvári Péter[2]

Becsült olvasási idő: 6 perc

Link: http://doi.org/10.13140/RG.2.2.14497.11360

Az AI különböző kutatási ciklusokra gyakorolt potenciális előnyeit jelenleg intenzív vizsgálatok övezik. Ebben a tanulmányban Kayvan Kousha és Mike Thelwall azt vizsgálják [1], hogy nagy nyelvi modellek, mint például a ChatGPT, alkalmazhatók-e a REF hatástanulmányainak minőségi értékelésére.

Az Egyesült Királyság Kutatási Kiválósági Keretrendszerének (REF) egyetemi értékeléseiben a hatástanulmányok ötoldalas, bizonyítékokon alapuló dokumentumok, amelyek azt mutatják be, hogy az adott egység (például egy tanszék) kutatásai milyen társadalmi hatásokat generáltak.

Egy hatástanulmány értékelése lényegesen nehezebb, mint a hagyományos tudományos outputoké, mivel rendkívül változatosak. Az értékelőknek figyelembe kell venniük a hatás eléréséhez vezető út erősségét, a hatás kiterjedését (szélességét és mélységét), valamint azt, hogy a hatásért mennyi elismerés illeti az adott tanszéket.

Ebben az összefüggésben hasznos lehet egy olyan eszköz, amely becsléseket ad a hatástanulmányok pontszámaira, hogy támogassa a döntéshozatali folyamatokat, különösen az egyetemek és akadémikusok számára, akiknek el kell dönteniük, mely tanulmányokat válasszák ki és hogyan készítsék el azokat. Végső soron ez egy szövegfeldolgozási feladat, így a nagy nyelvi modellek (LLM-ek), például a ChatGPT számára is megvalósítható. Ráadásul ezek a modellek már bizonyították, hogy képesek előre jelezni a REF minőségét tudományos cikkek esetében. [2] Ennek megfelelően letesztelték, és azt tapasztalták, hogy a ChatGPT képes elfogadhatóan előre jelezni a hatástanulmányok várható pontszámait, de diszciplínák közötti eltérések mellett, ahogy azt alább kifejtésre kerül.

Mi található egy hatástanulmányban?

A hatástanulmányok az alábbi szerkezetet követik:

- **Hatás összefoglalása (100 szó):** Rövid leírás az adott hatásról, amelyet a tanulmányok ismertetnek.
- **A kutatás alapjai (500 szó):** A hatáshoz vezető kutatások részletes leírása.

---

[1] Az alábbi közlemény a Nemzeti Közszolgálati Egyetem Államtudományi és Nemzetközi Tanulmányok Kar gondozásában megjelenő Államtudományi Hírlevél Tudományos sarok rovatában jelent meg. A hírlevélre az alábbi linken keresztül lehet jelentkezni: https://antk-dl.uni-nke.hu/at-newsletter-confirm/confirm/feliratkozas.php
Az oktatási anyagnak szánt tanulmány **What happens when you let ChatGPT assess impact case studies? https://blogs.lse.ac.uk/impactofsocialsciences/2025/01/15/what-happens-when-you-let-chatgpt-assess-impact-case-studies/** alapján készült.

[2] Egyetemi docens, Nemzeti Közszolgálati Egyetem, Államtudományi és Nemzetközi Tanulmányok Kar, Közszervezési és Infotechnológiai Tanszék, 1083 Budapest, Ludovika tér 2. E-mail: Sasvari.Peter@uni-nke.hu

- **Hivatkozások a kutatásra (6 hivatkozás):** Azoknak a kulcsfontosságú kutatási eredményeknek a hivatkozása, amelyek megalapozták a hatásokat.
- **A hatás részletei (750 szó):** Részletes beszámolók a hatásokról, beleértve, hogyan járultak hozzá a kutatások, és kik voltak az érintettek.
- **Források a hatás igazolására (10 hivatkozás):** A hatásokra vonatkozó állításokat alátámasztó bizonyítékok, például ajánlások vagy hivatalos jelentések.

A legtöbb hatástanulmány online is elérhető. [3]

Hogyan értékelhetik ezeket a ChatGPT segítségével?

Miután megvizsgálták, hogy a hatástanulmányok beküldése jogszerű-e szerzői jogi szempontból, a ChatGPT API-t használták az értékeléshez. Az API nem tanul a beadott adatokból. Ez tartalmaz egy rendszerpromptot, amelyet a feladat leírására lehet használni, valamint egy csevegési szekciót, amely egy kérdésből és a ChatGPT válaszából áll.

A rendszerpromptban a REF2021 hivatalos irányelveiből származó hatás- és minőségdefiníciókat, valamint a bírálóknak szóló panelkritériumokat és munkamódszerek dokumentumát használták fel. Ezeket úgy fogalmazták át, hogy önálló leírást nyújtsanak a hatásokról és azok értékeléséről. Az utasításokat az OpenAI példáinak stílusához igazították, amelyek főként azt írják le, hogy a ChatGPT-nek mit kell „*eljátszania*," nem pedig absztrakt módon mutatják be a feladatot.

Az A4-es oldal háromnegyedét kitevő rendszerutasítások [4] így kezdődtek:

„Önök tudományos szakértők, akik hatástanulmányokat értékelnek. Ezek olyan konkrét hatásokat ismertetnek, amelyek tudományos kutatásokból származnak. Önök 1* és 4* közötti pontszámot adnak, részletes indoklással együtt. ..."

Az eredményekhez csak az első rendszerutasítást használták, és nem próbáltak ki variációkat, mivel korábbi tapasztalataik azt mutatták, hogy ezeknek csekély hatása van. [5] Ezt követően a következő kérdést tették fel a ChatGPT-nek:

„Értékelje a következő hatástanulmányt:", majd megadták a címet, illetve a tanulmány egészét vagy részletét (lásd lentebb).

A **ChatGPT 4o-mini**[3] változatát használták az API-n keresztül. Minden tanulmányt ötször küldtek be értékelésre, és az átlagolt pontszámokat használták előrejelzésként.

Kísérletek és eredmények

A fent leírt megközelítést alkalmazták a 6220 nyilvánosan elérhető, megfelelő hatástanulmányra. Mivel a fő cél az értékelési pontszámok előrejelzése volt, nem pedig szakértői vélemények nyújtása az erősségekről és gyengeségekről, a legjobb előrejelzések elérésére összpontosítottak. Mivel nem ismert, hogy bármelyik hatástanulmány milyen pontszámot kapott, a REF2021 eredmények honlapján elérhető tanszéki átlagpontszámokat használták proxyként az egyéni pontszámokhoz. A cél az volt, hogy olyan előrejelzéseket kapjanak, amelyek a legnagyobb korrelációt mutatják ezekkel a tanszéki átlagokkal, nem pedig olyanokat, amelyek a legközelebb állnak hozzájuk. Ez azért volt fontos, mert a ChatGPT ebben a típusú feladatban sokkal jobb a

[3] A **ChatGPT 4o-mini** egy konkrét elnevezés az OpenAI egyik modellverziójára vagy konfigurációjára. Ez utal egy kisebb méretű vagy optimalizált verzióra, amelyet egy adott célra (pl. szövegek gyors feldolgozása) fejlesztettek.

pontszámok helyes sorrendjének meghatározásában, mint azok pontos értékeinek megállapításában. Magas korreláció esetén a ChatGPT által adott pontszámok könnyen skálára igazíthatók egy átalakítással vagy egy keresőtábla segítségével.

A teljes tanulmány vagy annak egyes részeinek bevitelét is kipróbálták. Az eredmények szerint a cím és az összefoglaló önmagában történő bevitele sokkal jobb előrejelzéseket adott (magasabb korrelációt a tanszéki átlagokkal), mint a teljes hatástanulmány bevitele. Érdekes módon a ChatGPT annyira „*lenyűgözőnek*” tűnt a teljes szövegek által, hogy gyakorlatilag mindegyiket 4*-osra értékelte!

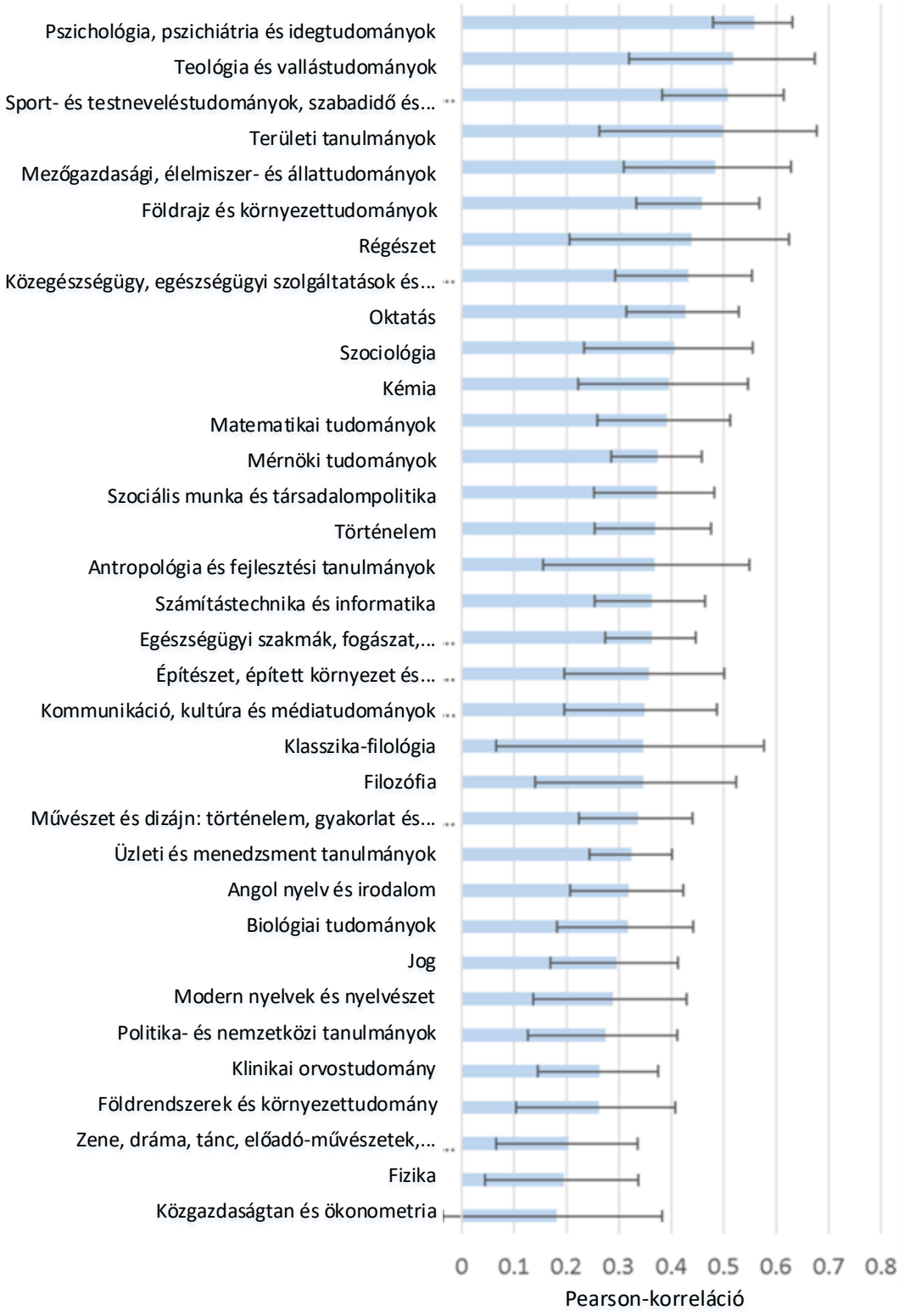
Az eredmények azt mutatják, hogy a ChatGPT-nek van valós, bár gyenge képessége a hatástanulmányok minőségének érzékelésére, amit a közvetett Pearson-korreláció (0,337) tükröz. Úgy tűnik, hogy leginkább az összefoglaló állítások „*megbízhatóságára*” támaszkodik, feltehetően azért, mert nem képes hatékonyan értékelni a részletes narratívát és az azokat alátámasztó bizonyítékokat.

Diszciplináris különbségek és promptok hatása

A kísérleteket megismételték szigorúbb promptokkal, hogy mérsékeljék a ChatGPT „*túlzott lelkesedését*” a 4*-os értékelések iránt, de ez csak csekély javulást eredményezett. A kísérletek során összehasonlították az egyes értékelési egységek eredményeit is, hogy illusztrálják, mely területeken működött a ChatGPT a legjobban és a legrosszabbul (1. ábra).

Ahogy a fentiek is sugallják, ha generatív mesterséges intelligenciát szeretne alkalmazni saját hatástanulmányainak pontozására, ezt megteheti a **ChatGPT API-val, de valóban csak az összefoglaló értékelésére érdemes használni, nem pedig a teljes dokumentumra**. A **folyamatot legalább ötször (de még jobb, ha harmincszor) meg kell ismételni**, és az átlagot kell figyelembe venni. Ez az átlag valószínűleg közel lesz a 4*-hoz, ezért az értéket érdemes figyelmen kívül hagyni, viszont használhatja arra, hogy összehasonlítsa az eredményt más, ugyanazon értékelési egységből származó hatástanulmányok pontszámaival.

Ez azonban csak egy ügyes tipp lesz, nem pedig egy megfelelő értékelés, ezért ne használja tényleges döntésekhez, hacsak nem merítette ki az összes ésszerű alternatívát!



1. ábra: A Pearson-korrelációk az egyes hatástanulmányok ChatGPT-átlagpontszáma és a tanszéki pontszám között, értékelési egységenként, 30 iteráció átlagát alapul véve, nagyon szigorú prompt alkalmazásával. Az értékelés során a hatástanulmányok címét és összefoglalóját vették figyelembe, félpontos skálázással. A hibasávok a 95%-os megbízhatósági intervallumokat jelzik.

Felhasznált irodalom

[1.] Kayvan Kousha, Mike Thelwall (2025): What happens when you let ChatGPT assess impact case studies? https://blogs.lse.ac.uk/impactofsocialsciences/2025/01/15/what-happens-when-you-let-chatgpt-assess-impact-case-studies/

[2.] Mike Thelwall, Abdallah Yaghi (2024): In which fields can ChatGPT detect journal article quality? An evaluation of REF2021 results, https://doi.org/10.48550/arXiv.2409.16695

[3.] Publications and reports, Latest publications from REF 2021 are listed by year in date order., https://2021.ref.ac.uk/publications-and-reports/index.html

[4.] Results and submissions, Introduction to the REF results, https://results2021.ref.ac.uk/

[5.] Kayvan Kousha, Mike Thelwall (2024): Assessing the societal influence of academic research with ChatGPT: Impact case study evaluations, https://doi.org/10.48550/arXiv.2410.19948