

# MAGYAR NYELVRE OPTIMALIZÁLT NYELVMODELLEK ALKALMAZÁSA KÖZGYŰJTEMÉNYI KÖRNYEZETBEN: LEHETŐSÉGEK ÉS AGGÁLYOK<sup>1</sup>

DOI 10.35402/kek.2026.1.12

## Absztrakt

A tanulmány a közgyűjtemények egy olyan jövő-képét vázolja fel, ahol a magyar igényekre szabott nagy nyelvi modellek a gyűjteményezés, a dokumentumok szövegének felismertetése, és a kutatástámogatás alapját képezik. A McLuhantól Castellsig ívelő médiaelméleti gondolkodás és a generatív MI-forradalom metszéspontjában új, „gutenbergi pillanat” bontakozik ki: a multimodális LLM-ek képesek gazdagítani a hiányos metaadatokat, átírni 19. századi kéziratokat, és keresővel támogatott chatbotokkal szolgálni a kiállítások látogatóinak igényeit. A szerző ugyanakkor nem hallgatja el a „rejtett költségeket” és veszélyeket sem, és konkrét magyar megoldási lehetőségeket és jó gyakorlatokat ismertet: nyílt, magas minőségű nemzeti korpuszt; a szuperszámítógépek drága és komplex használatát kiváltó eljárásokat; valamint olyan díjnyertes pilótákat, mint az Arany János akadémiai iratainak MI alapú feldolgozása, amely 30 000 oldalnyi dokumentum szövegét teszik kereshetővé.

*Kulcsszavak:* Nagy nyelvmODELLEK, digitális kulturális örökség, chatbot, kézírásfelismerés

## Abstract

This essay invites memory institutions to envision a future in which Hungarian-tailored large language models (LLMs) play a central role in cataloguing, text recognition, and research assistance. It blends classic media theory (from McLuhan to Castells) with today's generative-AI boom to argue that a new “Gutenberg moment” is unfolding in cultural heritage: multimodal LLMs can already enrich sparse metadata, transcribe 19th-century manuscripts and power retrieval-augmented chatbots for exhibition visitors. Yet the study also spotlights the hidden costs and presents concrete Hungarian solutions: an open, high-quality national corpus; lightweight fine-tuning

strategies that lower the super-computer barrier; and award-winning pilots that turn 30 000 pages of Arany János's Academic documents into searchable data.

*Keyword:* Large language models, digital cultural heritage, chatbot, handwritten text recognition

Jelen tanulmányt egy rövid elméleti bevezetével kezdem, Marshal McLuhan, Eric Havelock és Manuel Castells komplex médiaelméleti érveléséből villantok fel egy-egy gondolatot, ezután, Kate Crawford kritikai belátásait felidézve, a generatív mesterséges intelligencia néhány jellemzőjére térek ki, majd, röviden, bemutatok a digitális kulturális örökség területén zajló néhány jelenleg is folyó fejlesztést, amelyek ebben a kontextusban érdemlegesen lehetnek.

*Egy absztrakt festmény kreatív gondolati folyamata közvetlen reprezentációja, ahogyan azok a számítógép dizájnban is megjelenhetnek. Amiről azonban itt szó van, az a dizájnban vagy a mintázatoknak a pszichés és társadalmi következményei, ahogy azok felerosítik vagy felgyorsítják a meglévő folyamatokat. Hiszen bármely médium vagy technológia „üzenete” az a változás, a tempó vagy a mintázat vagy a lépték változása, amelyet az emberi dolgokba bevezet.<sup>2</sup>*

Nem pusztán arról van szó, hogy a médium volna az üzenet lényege, *the medium is the message*, hogy a közhellyé koptatott idézetet citáljam, de egyenesen arról, hogy a médium átgúr (*the medium is the message*): a technológiai médium átalakítja az ember önmagához és környezetéhez való viszonyát, a kommunikációt, a psziché és a társadalom szerkezetét.<sup>3</sup> Mindezt McLuhan az 1960-as években, vagyis a számítógépek létrejötte után, de a számítógépek közvetítette kommunikáció elterjedése előtt évtizedekkel fogalmazta meg, mintegy fél évszázaddal a generatív mesterséges intelligencia forradalmát megelőzve.

<sup>2</sup> Marshall McLuhan: *The Medium Is the Message: Understanding Media, The Extensions of Man*, 9., McGraw-Hill, New York, 1964.

<sup>3</sup> Marshall McLuhan–Quentin Fiore: *Médiamasszázs, Egy rakás hatás*, Typotex, 2012.

<sup>1</sup> Jelen szöveg a MuseumDigit konferencián, a Nemzeti Múzeumban 2024 novemberében elhangzott előadás szerkesztett, bővített változata.

Maradjunk még az 1960-as években egy pillanatra. Eric Havelock az „alfabetikus elme” fogalmának megalkotásával, párhuzamosan McLuhan érvelésével, arra a változásra hívja fel a figyelmet, hogy a lineáris betűírás hogyan alakította át a görög társadalmat, és végül, hogyan tette lehetővé a logika, és általában a tudomány és a műveltség létrejöttét azáltal, hogy a „szóbeli elme”, az írás által formált kultúra előtti emberi gondolkodás átalakul.

*A tapasztalat narratív megformálása nem az ókori nyelvek sajátos kifejezőmódja vagy különösége volt. Ez minden szóban megőrzött beszéd lényegi összetevője a világ minden nyelvén. A görög ábécé megjelent, átvette ezt, és újraformálta, létrehozva a nyelv és az elme új univerzumát; egy olyan univerzumot, amely alapelvekből, kapcsolatokból, törvényekből, és tudományokból, értékekből, eszmékből és ideálokból áll. Ezek most a közvetlen érzékszervi apparátusunk és az ezt az apparátust legkönnyebben kifejező szóbeliség fölött uralkodik. Egy vizuális nyelvi architektúra került a nyugtalan akusztikus hangáram fölé. Jó vagy rossz: ez az írástudás forradalmának gyümölcse a Nyugat számára.<sup>4</sup>*

És most ugorjunk egy nagyot az időben, a máig, vagy legalábbis a közelmúltig. Manuel Castells, a hálózati társadalom teoretikusa az alfabetikus írás megjelenéséhez hasonló nagyságrendű média- és kultúrtörténeti paradigmaváltásként állítja elének a számítógéppel közvetített, multimediális kommunikáció és az elektronikus feldolgozott információs hálózatok megjelenését, pontosabban ezek általános elterjedését. A hálózati társadalom 2700 évvel az alfabetikus elme megszületése után jön létre, radikálisan átalakítva a tér és az idő szerveződését, a személyes és a társas kapcsolatokat, vagy épp a politikai diskurzusok működését.

*Egy hasonló történelmi jelentőségű technológiai átalakulás zajlik 2700 évvel később, nevezetesen a kommunikáció különböző módjainak interaktív hálózatba való integrációja. Másképpen fogalmazva, egy hipertext és egy metanyelv kialakulása történik, amely először a történelemben egyesíti az írott, szóbeli és audiovizuális emberi kommunikációs módokat ugyanazon rendszerben.<sup>5</sup>*

A médiaevolúció fenti, kritikai belátásai fölől tekintve jól látható, hogy az olyan hibrid,

a ma kommunikációját meghatározó kultúrtechnológiák, mint például a *social media* platformok, milyen mélyreható változásokkal járnak a társadalom működésének egészére, a kultúra előállítására és befogadására. Castells egy szép esszében elemzi, a valós virtualitás korában hogyan alakul át a múzeumok szerepe.<sup>6</sup> Az azonban egyelőre nehezen megjósolható, hogy a generatív MI minden képzeletet felülmúlóan gyors térnyerése vajon hasonlóan átható következményekkel jár-e. Az olyan kritikus hangok, mint Kate Crawfordé, az *Atlas of AI* című kötet szerzőjéé, amely hangok az MI technológia etikai és szociális implikációira mutatnak rá, a generatív MI penetrációjának vakító fényében (hogy képzavarral éljek), nagyon is halknak bizonyulhatnak.<sup>7</sup> (Itt, kissé pongyolán, McLuhan a látványra és a hallásra épülő kultúrák ellentétére utalok.) Ha a szöveggenerálás, az alfabetikus elme privilégiuma, a kultúraalkotás, mint tevékenység és felelősség, mindinkább a mesterséges intelligencia kezébe kerül, annak következményei valóban beláthatatlanok. A sorsnak azonban bizonyosan van, kissé tán keserű, humorérzéke, az „alfabetikus elme”, amelynek kiteljesedése a *Gutenberg galaxis* kialakulásához kötődik,<sup>8</sup> talán éppen azon a ponton éli meg tudástermelő privilégiumának elvesztését, amikor a nagy nyelvmodellek emergens képességei megjelennek,<sup>9</sup> vagyis amikor milliárd oldalnyi ember alkotta, írott szöveg algoritmikus feldolgozásával megjelenik a generatív MI, a Turing tesztet nevetve teljesítő, kreatív, író, beszélő, beszélgető, sőt a fenti képességekre építve már képeket és zenét alkotó gép.

De mit is jelent a generatív MI alapját képező nagy nyelvmodellek, az LLM-ek emergens képessége, amely a ChatGPT nevű szolgáltatás révén szakította át az MI nyarak és telek<sup>10</sup> évtizedes váltakozásába belefáradt médiafigyelmet? Vektorszemantikára épülő nyelvmodellek, például a

<sup>6</sup> Manuel Castells: *Múzeumok az információtól a virtuális múzeumig* In.: *Múzeumelmélet (A képzeletbeli múzeumtól a hálózati múzeumig)*, Szerk.: Palkó Gábor, Ráció, 2012.

<sup>7</sup> Kate Crawford: *Atlas of AI. Power, Politics, and the Planetary Costs of Artificial Intelligence*, Yale UP, 2021.

<sup>8</sup> Marshall McLuhan: *The Gutenberg Galaxy. The Making of Typographic Man*. University of Toronto Press, 1962. Magyarul: *A Gutenberg-galaxis. A tipográfiai ember létrejötte*. Trezor Kiadó, 2001.

<sup>9</sup> Wei, Jason, et al. „Emergent abilities of large language models.” arXiv preprint arXiv:2206.07682 (2022).

<sup>10</sup> Hendler, James. „Avoiding another AI winter.” *IEEE Intelligent Systems* 23.02 (2008): 2–4.

<sup>4</sup> Eric A. Havelock: *The Alphabetic Mind: A Gift of Greece to the Modern World* In: *Oral Tradition*, 1/1 (1986), 149.

<sup>5</sup> Manuel Castells: *The Rise of the Network Society*, Wiley-Blackwell, 210, 356.

Word2Vec<sup>11</sup> technológia, már jóval hamarabb léteztek, és rengeteg olyan feladatot képesek voltak magas színvonalon elvégezni, amelyek korábban fáradságos emberi közreműködéssel valósulhattak meg, mint például a szövegklasszifikáció, a nyelvedetektálás vagy az OCR-ezett szövegek minőségének ellenőrzése és akár javítása. Ezeket a modelleket azonban többnyire az adott nyelv adott végrehajtandó feladatára kellett hangolni, vagyis a létrejött modellek többnyire egyetlen feladat végrehajtására voltak csak használhatók. Az, hogy rövid, emberi nyelven megfogalmazott prompt, esetleg néhány mintapélda alapján a modell által sohasem tanult feladatok is elvégezhetővé válnak a modell módosítása nélkül, tulajdonképpen ez a váratlan minőségi ugrás hozta el az MI, mint technomédiium jelenleg is dübörgő forradalmát.<sup>12</sup>

De hogyan érinti mindez a Gutenberg-galaxis örzöit, a memóriaintézményeket? Három fogalmat szeretnék kiemelni a gyűjtemények lehetőségei és felelőssége kapcsán: a korpusz, az elfogultság és a fekete doboz első hallásra talán össze nem függő fogalmait.

Kezdjük az elfogultság, vagy az angol szakszóval: a *bias* jelenségével.<sup>13</sup> Ahogyan Kate Crawford hangsúlyozza, az elfogultság az MI területén nem pusztán technikai probléma vagy programhiba, sokkal inkább a tanítóanyagokban benne rejlő társadalmi, történelmi, intézményes elfogultságok akkumulációja, amelynek súlyos társadalmi következményei lehetnek.<sup>14</sup> Bármilyen furcsán is hangzik, ez az elfogultság kvantifikálható, mérhető, legalábbis mérnöki értelemben, hiszen a nyelvmodellek inherens jellemzőjéről van szó, és a modellek összemérhetősége szempontjából ez elengedhetetlen. Az első képen a Meta Llama modelljének elfogultság-értékei láthatók, más LLM-ekkel összehasonlítva. Jól látható, hogy az összevetett modellek mindegyike szexista,

rasszista, szociális, vallási és egyéb elfogultságok megtestesítője.<sup>15</sup>

	LLaMA	GPT3	OPT
Gender	70.6	<b>62.6</b>	65.7
Religion	79.0	73.3	<b>68.6</b>
Race/Color	<b>57.0</b>	64.7	68.6
Sexual orientation	81.0	<b>76.2</b>	78.6
Age	70.1	<b>64.4</b>	67.8
Nationality	64.2	<b>61.6</b>	62.9
Disability	<b>66.7</b>	76.7	76.7
Physical appearance	77.8	<b>74.6</b>	76.2
Socioeconomic status	<b>71.5</b>	73.8	76.2
Average	<b>66.6</b>	67.2	69.5

Table 12: **CrowS-Pairs.** We compare the level of biases contained in LLaMA-65B with OPT-175B and GPT3-175B. Higher score indicates higher bias.

Miért ennyire elfogultak a modellek? És miért érzékelhető ez egyre kevésbé az olyan érett szolgáltatásokban ma, mint a ChatGPT vagy a Microsoft Copilot, amelyek elfogult LLM-ekre épülnek? A válasz az LLM-ekre épülő szolgáltatások kialakításából adódik. A nyers modellek gigantikus szövegtörzseken tanulnak, ezután instrukciókat tartalmazó tanítóanyaggal megtanítják a már nyelvi képességekkel rendelkező modelleket utasításokat követni, majd megtanítják őket viselkedni a megerősítéses tanulás technológiájával. Már ezen a ponton is felmerül a fekete doboz effektusa, hiszen a korpuszok összetétele, a szövegek pontos forrása többnyire még a közzétett és saját hardware-környezetben is használható „nyílt” modellek esetén is ismeretlen. Az elfogultság kezelésére azonban a fent említett, multinacionális óriáscégek kínálta szolgáltatások további, üzleti titokként kezelt technológiákat integrálnak az elfogultság ellensúlyozására, a sértő vagy nem PC tartalmak generálásának megakadályozására, ezek szűrésére, ráadásul ezeket folyamatosan, akár napi szinten módosítják. Van abban valami rémisztő, hogy ismeretlen szövegtörzseken és instrukciókkal tanított, bonyolult cenzurális mechanizmusokat integráló, naponta változó generatív MI eszközök épülnek be a memóriaintézmények napi gyakorlatába, és ebben az értelemben a digitális kulturális örökség cirkulációjába. Márpedig ez nehezen látszik elkerülhetőnek, egy olyan szituációban, amikor a

<sup>11</sup> Mikolov, Tomas, et al. „Distributed representations of words and phrases and their compositionality.” *Advances in neural information processing systems* 26 (2013). A nyelvmodellek fejlődéséhez: Mars, Mourad. „From word embeddings to pre-trained language models: A state-of-the-art walkthrough.” *Applied Sciences* 12.17 (2022): 8805.

<sup>12</sup> Brown, Tom, et al. „Language models are few-shot learners.” *Advances in neural information processing systems* 33 (2020): 1877–1901.

<sup>13</sup> Gallegos, Isabel O., et al. „Bias and fairness in large language models: A survey.” *Computational Linguistics* (2024): 1–79.

<sup>14</sup> Crawford, 128. skk.

<sup>15</sup> Touvron, Hugo, et al. „Llama: Open and efficient foundation language models.” *arXiv preprint arXiv:2302.13971* (2023).

kulturális örökség objektumok többségéről kevés és többnyire nem szemantikus metaadat áll rendelkezésre, gondoljunk csak az Europeana szolgáltatás 50 millió rekordjára, miközben a multimodális LLM-ekre épülő szolgáltatások egyre komplexebb és egyre minőségibb tárgyleírások készítésére képesek, akár különösebb előtanítás nélkül is.

Nem kétséges, hogy a kulturális örökség intézmények és a digitális tudományosság, az eScience elemi érdeke, hogy nyílt, a fekete doboz effektusától mentes vagy kevésbé érintett modellek legyenek a tudományos és kulturális munkafolyamatokba illeszthetők,<sup>16</sup>

biztosítva, hogy a kutatások és a munkamenetek FAIR alapelveknek megfelelően átláthatók, megismételhetők, megoszthatók legyenek.<sup>17</sup> Ez azonban nem triviális, mivel a beépési küszöb a generatív MI technológiák esetében több értelemben is rendkívül magas. Magas a fejlesztéshez szükséges IT és nyelvtchnológiai kompetenciák szempontjából, de a hatalmas emberi erőforrás és számítási igény miatt a fejlesztés és a napi használat költsége is rendkívül magas. Intézményközi összefogás nélkül az ilyen fejlesztések szinte elképzelhetetlenek, különösen egy erősen alulfinanszírozott kulturális és tudományos környezetben.

ChatGPT 4o



You are a Hungarian art historian. Please write a short description of the enclosed painting. The artist is Csontváry, Kosztká Tivadar, the title is "Önarckép".



The painting titled "**Önarckép**" (**Self-Portrait**) by **Kosztká Tivadar Csontváry** is a compelling and introspective work by one of Hungary's most renowned and visionary artists. This oil painting showcases Csontváry's remarkable ability to capture psychological depth through portraiture. The artist's penetrating gaze and stern expression dominate the composition, reflecting both his introspection and a sense of individuality. The use of a dark background emphasizes the figure, while the sharp features of the face, particularly the intense blue-green eyes and well-defined beard, suggest a meticulous attention to detail and emotional intensity. Csontváry, known for his innovative style and mysticism, conveys a unique narrative of self-awareness and artistic determination in this work. The simplicity of the attire, with a formal black suit and tie, contrasts with the complex emotional undertones, making this self-portrait a striking study of the artist's inner world.

🔊 📄 👍 🗨️ ↻

<sup>16</sup> A fekete doboz felnyitása („unblackboxing”) fogalmához médiatechnológiai perspektívából l. pl. Rieder, B., Peeters, S., & Borra, E. (2024). From tool to tool-making: Reflections on authorship in social media research software. *Convergence*, 30(1), 216-235. <https://doi.org/10.1177/13548565221127094>

Ugyanakkor ennek a fejlesztésnek, különösen Magyarországon, kiemelt jelentősége van, nem túlzás, nemzeti érdek. Nem véletlen, hogy

<sup>17</sup> A FAIR alapelvekről röviden: <https://openscience.hu/f-a-i-r-kutatasi-adatkezes/> (Hozzáférés: 2025.01.12.)

Svédországban egy jól finanszírozott, állami támogatással működő intézmény, az AI Sweden szervezi azt.<sup>18</sup> A magyar helyzet még a svédnél is kényesebb, hiszen, és itt térjünk vissza a korpuszok kérdéséhez, a generatív MI eszközök teljesítménye jelentős részben a tanító korpusz méretén és nyelvi minőségén múlik, és a kisebb digitálisan elérhető szövegmenyiség komoly hátrányt jelent.<sup>19</sup> A svéd példa azonban, hogy az LLM-ek fejlesztésébe a rokon nyelvek hasonló méretű korpuszait integrálják, a magyar nyelv esetében, nyilvánvalóan, nem áll rendelkezésre.

A magyar nyelvre optimalizált nyelvmodellek fejlesztésének nemzetpolitikai potenciálját felismerve indult el az akkori Innovációs és Technológiai Minisztérium és az OTP közreműködésével az a projekt, amelynek célja a fenti elvárásoknak megfelelő modellek létrehozása.<sup>20</sup>

A Digitális Örökség Nemzeti Laboratórium, és annak szakmai vezetője, az ELTE Digitális Bölcsészeti Tanszék felel a fenti projekt magyar nyelvű korpuszána létrehozásáért, és fő célunk, a korpusz hatalmas mérete és magas nyelvtechnológiai minősége mellett, a kulturális-tudományos szempontból értékes tartalmak arányának maximalizálása. Ez rámutat a memóriaintézmények felelősségére a jövő mesterséges intelligenciájának fejlesztésében, az értékes szöveges tartalmak (lehetőleg nem csak emberi olvasásra alkalmas) tömeges közzétételében. A ma széles körben használatos modellek ugyanis többségében az internet válogatatlan anyagára épülnek, és ez a többnyelvű LLM-ek magyar nyelvű alkorpuszára is igaz: „Országok rongya! könyvtár a neved”: „Internet szemete, nagy nyelvmódel a neved.”

Hogyan érinti mindez a digitális kulturális örökség szakembereket és fejlesztéseket? Három jelenleg is

zajló kísérletről szeretnék ezzel kapcsolatban, nagyon röviden, beszámolni.

A ELTE BTK Digitális Bölcsészeti Tanszéke és az ELTE Informatikai Kar Mesterséges Intelligencia Tanszéke olyan kísérleti fejlesztésbe kezdett, amely egy létező többnyelvű LLM átalakításával egy, a magyar nyelvre, ezen belül kulturális és tudományos szövegek feldolgozására alkalmas, de kisebb, vagyis százmilliós nagyságrendű befektetést igénylő szuperszámítógép használata nélkül is alkalmazható. Technológiai szempontból ez egy rendkívül komplex feladat, mint a kutatócsoport nyelvtechnológus-mérnök szakértői megállapították, a sikeres modellfejlesztés kulcsa az eredeti, nagyon kis magyar nyelvű korpuszon tanított LLM soknyelvű szótárának manipulálása lehet.<sup>21</sup>

Ami a tanításhoz, vagy pontosabban, továbbtanításhoz szükséges korpuszt illeti, itt egy – meglepő módon nem csak magyar nyelvű – tudományos szövegeket és primer szövegkiadásokat különféle dokumentum-formátumokban tartalmazó korpusz létrehozása volt a feladat. Miért többnyelvű, ha a cél a magyar nyelvi kompetenciák növelése? A tapasztalatok szerint az LLM-ek teljesítménye a továbbtanítás során jelentősen romlik, különösen a fordítási és programozási feladatok esetén, ha a tanítást egyetlen célnyelven folytatják. És miért a különféle dokumentumformátumok, mint az XML, a JSON, a LaTeX és az RDF? Azért, mert a kulturális örökség szemantikus feldolgozásának ezen szabványok és az ilyen formátumokat és szabványokat alkalmazó adatbázisok az alapjai. Arról, hogy ezek a kísérletek hova vezetnek, csak jóslataink és reményeink vannak.

A második példa ennél jóval konkrétabb. A 2024-ben a HUN-REN Bölcsészettudományi Kutatóközpont a DH-LAB keretei között folyó, Arany János hivatali iratainak MI alapú feldolgozása című projektjével Társadalmi Innovációs Díjat nyert a Kulturális és Innovációs Minisztériumtól.<sup>22</sup> A projektről már több előadáson és publikációban is beszámoltunk, például a Networkshop konferencián többször is, ezért csak

<sup>18</sup> AI Sweden: „Sweden’s national center for applied artificial intelligence” A szervezet szerepéről l. pl. The innovation intermediary’s guide to the AI Factory - A study of AI Sweden as an intermediary in a system of innovation <http://hdl.handle.net/2077/69125>

<sup>19</sup> „Minority languages are vital to preserving cultural heritage, yet they face growing risks of extinction due to limited digital resources and the dominance of artificial intelligence models trained on high-resource languages.” Carta, Salvatore Mario, et al. „LIMBA: An Open-Source Framework for the Preservation and Valorization of Low-Resource Languages using Generative Models.” arXiv preprint arXiv:2411.13453 (2024).

<sup>20</sup> Ld. az ezzel kapcsolatos bejelentéshez: <https://kormany.hu/hirek/az-itm-az-otp-vel-europa-legnagyobb-szuperszamitogepenek-fejlesztesebe-kezd> (Hozzáférés: 2025.01.12.)

<sup>21</sup> Az LLM-ek „szótára” valójában nem szótár a szó kulturális értelmében, hiszen az adott esetben százat is meghaladó többnyelvű korpusz leggyakoribb szó- és mondatöredékeiből áll. A kérdéshez l. pl. Kozma, László, and Johannes Voderholzer. „Theoretical Analysis of Byte-Pair Encoding.” arXiv preprint arXiv:2411.08671 (2024).

<sup>22</sup> <https://iti.abtk.hu/hu/osztalyok/digiphilcsoport/1778-tarsadalmi-innovacios-dij-akezirasfelismero-modellek-fejleszteseert> (Hozzáférés: 2025.01.12.)

röviden foglalom össze annak lényegét.<sup>23</sup> A kéziratok, amelyek a kulturális örökség felbecsülhetetlen értékű hányadát hordozzák, időigényes és nagy szakértelmet igénylő feldolgozásuk miatt a digitális kulturális örökség egészét tekintve súlyosan alulreprezentáltak. A gépi kézírásfelismerés vagy becenevén HTR technológia lehetővé teszi, hogy ez a folyamat radikálisan felgyorsulhasson, így a kéziratokban rejlő tudás nagyobb arányban kerülhessen be a tudományos diskurzusokba, az emberi és a gépi olvasás gyakorlataiba. A DHLAB, azért, hogy az ehhez szükséges technológia nagy korpuszok feldolgozására, gyűjteményi felhasználásra elérhetővé váljon, együttműködő partnereivel, mint a Nemzeti Levéltár és az MTA könyvtára, HTR modelleket fejleszt. Ennek az egyik első sikere az Arany János hivatali iratainak kereshetővé tétele 30.000 oldal terjedelemben. A modell az Irodalomtudományi Intézet és a Miskolci Egyetem irodalomtörténész szakértői által készített mintegy 1000 kéziratlap átírásán tanult, és 6% alatti betű szintű hibaarányt produkál a tucatnyi kéz által írott dokumentumokon. Ennél is jobb eredményt értünk el az OSZK Kulcsár Péter Kutatócsoportjával együttműködve latin nyelvű kéziratok átírásával, a legújabb projektünk pedig Pázmány Péter műveinek több ezer oldalra rúgó korpuszának feldolgozása. Itt korai nyomtatványokról van szó, a feladat valahol félúton található a HTR és az OCR technológia között, és az eredmények (legalábbis, ha a modellt egy-egy konkrét nyomda kiadásaira tanítjuk) biztatóak. A dokumentumok betűinek, szavainak és sorainak felismerése azonban csak az első lépés. Arra törekszünk, hogy LLM technológia segítségével (egyes tanulmányok szerint a multimodális, vagyis nem csak olvasó, de látó LLM-ek hamarosan leváltják a HTR és OCR céleszközöket), az Arany főútkári időszakához kötődő hatalmas iratmennyiséget metaadatokkal lássuk el. Meglepetők az első eredmények, az általunk jelenleg használt Llama 3.1 70B modell a felismertetés hibáit sok esetben korrigálva képes metaadatok kinyerésére és a 19. századi dokumentumok értelmezésére, szöveges összefoglalók készítésére. Személy- és földrajzi nevek, dátumok kinyerése mellett jó határfokkal ismeri fel a modellt a műfajokat és az iratok funkcióját. Mivel saját szerverten futtatott modelltől van szó, a nyelvi képességeik terén nehezen utolérhető ChatGPT borsos ára sem akadályozza a projekt végrehajtását.

<sup>23</sup> Palkó, Gábor and Szekrényes, István and Bobák, Barbara (2023) A Digitális Örökség Nemzeti Laboratórium webszolgáltatásai automatikus kézírás-felismertetéshez. In: Új technológiákkal, új tartalmakkal a jövő digitális transzformációja felé. HUNGARNET Egyesület, Budapest, pp. 164-169.

És végül egy, az eddigieknél is merészebb, folyamatban lévő fejlesztésről szeretnék beszámolni. Korábban nem említettem a generatív mesterséges intelligencia technológiai diskurzusából a közbeszédbe is átkerült, félrevezető terminussal hallucinációnak nevezett jelenséget, amely az LLM-ek felhasználásának fő akadálya kulturális örökséget érintő és tudományos feladatok megoldásában. Kézenfekvő megoldást kínál a problémára a rendkívül divatosá vált RAG technológia.<sup>24</sup> Egy egyszerű allegóriával világítanám meg ennek működését: ha az LLM országok rongyán, könyvtárak dokumentum-kilométerein tanult is, képességei absztraktak, a generált tartalmak távol állnak az egyes művekben reprezentált tudástól, illetve csak a tanítása során gyakran ismétlődő tudáselemekre emlékezik. Olyan, mint egy, kissé már feledékeny professzor, a könyvtárától elszakítva. A RAG technológia az LLM emergens képességeit és szakkönyvek tudását ötvözi, ezért egy adott témában sokkal pontosabban képes a feladatok megoldására. Gondoljunk bele ennek a jelentőségébe. Az Arany hivatali iratok 30.000 oldalát végigolvasni hatalmas időráfordítás, ráadásul egy-egy konkrét kutatói kérdés alapján az iratok más, esetenként csekély részhalmaza válik csak érdekessé, és ez a halmaz kérdésenként más és más. Még radikálisabb feladat, és a jelen pénzügyi környezetben kissé álomszerű, ám technológiailag megoldható volna egy olyan alkalmazás fejlesztése, amely, hogy ideillő példát hozzak, a Nemzeti Múzeum Régészeti Intézetének milliós ásatási dokumentációjából kinyert tudásra épülő kutatói asszisztens építése.

De maradjunk a realitásoknál. A két fent említett tanszék együttműködésében egy olyan MI asszisztens megalkotásán dolgozunk, amely a Magyar Tudományos Akadémia alapításának 200-ik évfordulójának alkalmából rendezett kiállításához kapcsolódva az Akadémia történetét dokumentáló tudományos források alapján válaszol az érdeklődők kérdéseire. Az, hogy mennyire ad tényszerűen pontos és értelmes, jól formált mondatokat, számtalan faktor összjátékán múlik. Az egyik, természetesen a korpusz összeállítása, erre az MTA történetét tanulmányozó kutatók vállalkoztak. De legalább ennyire fontos az alapmodell, az LLM kiválasztása.

És ezzel a kérdéssel zárom le ezt a rövid tanulmányt. Ha kulturális örökség feldolgozására, vagy kutatói célra használunk MI technológiát (a kettő nyilván összekapcsolódik), akkor vajon elég, ha valamelyik

<sup>24</sup> Lewis, Patrick, et al. „Retrieval-augmented generation for knowledge-intensive nlp tasks.” *Advances in Neural Information Processing Systems* 33 (2020): 9459–9474.

Not secure molybclick.elle-dh.hu/3000

EDGE DH Chatbot Running: burgdig@quantis:libeta.Llama3.3.1.70B-Instruct-AWD-INT4

Chatbot

This is a document written in Hungarian in the 19. Century. The text is created by a handwritten text recognition model, so there are layout errors and character errors. It is an official document of the Hungarian Academy of Sciences. Please find person names in the text. Answer with a list of person names. "Chyzer Kornél 207 Eperjes. Tekintetes Úri rencsém van tudatni, hogy értekezésért a tiszt. díj (Sáros megye tisztelét dy 1864 dec ásvány vizei 109 ft. m. 6-án utalványoz- tatott. Méltotassék tehát a Péntztárhoz, vagy az illető Éresítő szerkesztőjéhez for- dúlni Maradván tisztelettel Pest, Tekint. 65 dec. 16. 1868 MAGYAR aláz 7. észénő "

Now find place names, if any.

Now find dates and provide them in ISO format.

Based on the provided text, I found the following person name:

1. Chyzer Kornél

Based on the provided text, I found the following place names:

1. Eperjes
2. Pest
3. Sáros ( likely referring to Sáros county, a former administrative division in the Kingdom of Hungary, with its center in Eperjes, now Prešov in Slovakia)

Based on the provided text, I found the following dates in ISO format:

1. 1864-12-06 (December 6, 1864)

Here is another document. Please find dates (if any) and list them in ISO format. "kelt Abonyban 1865. (ezer nyolczszáz kilenczven ötödikjévi október hó 15 (tizenharmadik) napján-A lezárás és az üres helyek kitöltése után -az aláíráse előtt – kitűnt, hogy a számok szavakkal nem lettek kúrva. Ugyanazért beje lentő meglegzi hogy - I, a bejelentést ezer nyolczszáz kelenczven ötödik évi október tizen harmadikán tét. tért. -- II, az aya huszon nyolcz éves. III, az anya huszon nyolcz éves IV, a születés ezerny ol száz kilencz ven ötödik évi október tizen ötödik kán történt déi előtt két negyed tizen kétő- rak Gönczözi Pális. máthe-József s. k. anyakönyvvezető Az első példánnyal való: meggyerés hiteiélű Kelt Abonyban 1865. ok- töber hó 13-án MáthóSözset Aurak Kározos visszerintő. NB.79/1920 "

Try to make a summary of the document!

Here are the dates found in the document in ISO format:

1. 1865-10-15 (October 15, 1865)
2. 1865-10-13 (October 13, 1865)

The document appears to be a birth certificate or a record of birth, issued in Abony in 1865. The certificate contains information about the birth, including:

- o The date of birth: October 15, 1865
- o The time of birth: before noon, at 11:45
- o The ages of the parents: the father is 29 years old, and the mother is 28 years old

The document is signed by Gönczözi Pál and Mátyás József, who is the registry clerk (anyakönyvvezető). There is also a note at the end, dated 1920, but its relevance to the birth record is unclear.

multinacionális vállalat szolgáltatását használjuk, pénzt és emberi tudást dobálva annak egymásba skatulyázott fekete dobozaiba, reménykedve a jó minőségű feladatmegoldásban és népszerű alkalmazások működtetésében? Vagy járatlan utakra merészkedünk, vállalva ennek minden kockázatát, és a nyílt tudományosság jegyében megosztható tudást és átlátható eszközöket építünk? Döntse el mindenki magának. Bár a jelen magyar muzeológusai, könyvtárosai, a memóriaintézmények és a tudomány napszámossai számára ilyen döntések nemigen adódnak.

Sokkal valószínűbb, hogy ezeket a piacvezető szoftverfejlesztő cégek vezetői és marketingesei hozzák meg, és, akár a közműszolgáltatók weboldalain, egyszer csak csetbotok jelennek meg a gyűjteménykezelő rendszerek vagy a tudományos forrásgyűjtemények grafikus felületeinek sarkában, hogy „asszisztálják”, idézőjelben, a tudástermelést.<sup>25</sup> De hogy azok az LLM-ek, amelyek ezen szolgáltatások mélyén, saját elfogultságaik tengerében és öncenzúrájuk hálójában tárgyleírásokat és tudományos cikkek összefoglalóit írják, bizonyosan láthatatlanok maradnak.

---

<sup>25</sup> Az Arcanum pl. saját modellt készített, hogy a digitális gyűjtemény szövegei alapján az MI emberi kérdéseket tudjon megválaszolni, de sem a tanítóanyag, sem a modell nem nyilvános, így valódi fekete dobozként funkcionál. A cikk lezárásának időpontjában (2025.01.12.) a demo felület (<https://www.arcanum.com/hu/arcanum-ai/>) nem működött. Egészen más filozófia áll a Gale Digital Scholar Lab mögött: a források feldolgozásához digitális bölcsészeti eszközöket kínálnak a felhasználóknak. Ld. <https://www.gale.com/intl/primary-sources/digital-scholar-lab> (Hozzáférés: 2025.01.12.)