# A TUDOMÁNY, AZ OKTATÁS ÉS A KÖZGYŰJTEMÉNYI KISZOLGÁLÁS ÚJ INFORMATIKAI SZINERGIÁI

**NETWORKSHOP 2026**
**35. Országos Informatikai Konferencia**

**2026. március 31–április 2.**
**Debreceni Egyetem, Debrecen**

**Szerkesztette: Tick József, Kokas Károly, Holl András**

**HUNGARNET Egyesület**
**Budapest, 2026**


HUN-REN
Magyar Kutatási Hálózat

# NETWORKSHOP

Szerkesztette: Tick József, Kokas Károly, Holl András

Tipográfia és tördelés: Vas Viktória
Korrektúra: Danyi Melinda
Angol nyelvi lektor: Lukács Katalin

Networkshop 2026 konferencia előadásainak közleményei
Debreceni Egyetem, Debrecen
2026. március 31–április 2.

ISBN 978-615-6792-29-7
DOI: https://doi.org/10.31915/NWS.2026

Kiadja a HUNGARNET Egyesület
az MTA Könyvtár és Információs Központ közreműködésével
Budapest
2026

Borítókép: freepik.com

DOI: 10.31915/NWS.2026.5



# A RAG A VÉGFELHASZNÁLÓI SZINTEN

# RAG AT THE END-USER LEVEL

*Ungváry Rudolf*
*Országos Széchényi Könyvtár*
*ungvaryr@gmail.com*

## Absztrakt

A RAG (Retrieval-Augmented Generation)– a *forráselőírt információkeresés* – elfogadott magyar meghatározása még nem létezik. Ami van, azt lényegében az informatika szaknyelvén fogalmazzák meg. (Pl. beágyazott visszakereséssel végzett generálás, lekéréses kibővített generáció, **a generatív MI képességeit külső információforrásokkal ötvöző adatlekérdezés,** visszakereséssel bővített generálás, adatlekérésre alapozott generálás/ mesterséges intelligenciák.) A természetes nyelvet használó, nem informatikus végfelhasználó számára nehezen derül ki, hogy pontosan, az ő gondolkodása, az ő teendői szempontjából miről van szó. Egyszerűen fogalmazva meg kell adnia általa kiválasztott forrásokat, beleértve külső adatbázisokat is az MI rendszerének, (mint amilyen pl. a ChatGPT), hogy azokat is „beágyazva" a rendszer elvégezze a megadott tárgyú információ keresését. A többnyire speciális, olykor az interneten hiányzó forrásokat megadva a felhasználó pontosabb, szakszerűbb válaszokat kap, adott esetben a saját szakterületén használatos különleges szakkifejezésekkel megfogalmazva. Más szóval a generatív MI nem csak saját „tudására", hanem konkrét, külső forrásokra is támaszkodik a válaszadáskor. Kétségtelen, hogy ezáltal a felhasználó jobb eredményeket kaphat, mintha e források nélkül kérdezne (fogalmazná meg a promtot). A 2025. évi konferencián már ismertettünk ilyen külső, a felhasználó által megadott forrást alkalmazó eljárást a MARC21 magyarra fordításához, de magával az MI-vel végzett, forráselőírt információkereséssel nem foglalkoztunk. Most ismertetjük, hogyan használhatja a végfelhasználó a számára rendelkezésre álló eljárást, és összehasonlítjuk a RAG ama fejlesztői, ipari, professzionális változatával, melynek kialakításához informatikai szakismeret szükséges – noha használatához, ha jó a kialakított rendszer, a felhasználónak szintén nincs szüksége beható informatikai szakismeretekre. Konklúziók: 1. Az LLM nem tudástár. 2. „Tudásához" meg kell adni a forrásokat. 3. A legspecifikusabb információk nincsenek a hálón 4 Nem a legfrissebb információ van a hálón. 5. Az igazán összetett, mélységi forráselőírt kereséshez nem elég a végfelhasználói ismeret. 6. A jövő a specializálódott RAG szövegalapú párbeszédrendszereké („chatbot").

## Abstract

The accepted Hungarian definition of *RAG* – source-prescribed information retrieval – does not yet exist. What does exist is essentially formulated in the technical language of computer science (e.g., generation with embedded retrieval, retrieval-augmented generation, data querying that combines the capabilities of generative AI with external information sources, generation enhanced with retrieval, generation based on data retrieval/artificial intelligences). For an end user who is not an IT specialist but uses natural language, it is difficult to understand exactly what this means from the perspective of their own thinking and tasks.

Simply put, the user must provide their AI system (such as ChatGPT) with selected sources – including external databases – so that the system can incorporate ("embed") them and perform information retrieval on the specified topic. By supplying sources that are often specialized and sometimes absent from the internet, the user can obtain more precise and professionally formulated answers, in some cases expressed using the special terminology of their own field. In other words, generative AI relies not only on its own "knowledge" but also on specific external sources when producing answers. There is no doubt that in this way the user can obtain better results than by asking questions (formulated prompts) without such sources. At the 2025 conference we already presented a procedure using such external, user-provided sources for the translation of MARC21 into Hungarian, but we did not address source-prescribed information retrieval carried out with AI itself. We now describe how the end user can employ the method available to them, and we compare it with the developer-level, industrial, professional version of RAG, the creation of which requires IT expertise – although, if the system is well designed, its use likewise does not require deep IT knowledge from the user. Conclusions: 1. An LLM is not a knowledge repository. 2. Sources must be provided for its "knowledge.". 3. The most specific information is not on the web. 5. Truly complex, in-depth source-prescribed retrieval cannot be carried out with end-user knowledge alone. 6. The future belongs to specialized, RAG-based text dialogue systems ("chatbots")

## Bevezető

A 2025. évi NETWORKSHOP konferencián már ismertettünk egy RAG-eljárást a MARC21 magyarra fordításához (Ungváry, 2025). Ebben magával a nagy nyelvi modelleken alapuló mesterséges intelligencia rendszerrel (LLM, a továbbiakban MI-vel) végzett, angol nevén Retrieval-Augmented Generation (RAG) eljárásával, annak magyar megevezésével és alkalmazásának formáival nem foglalkoztunk. Ennek a munkának ez lesz a tárgya.

## Terminológia

A RAG elfogadott magyar meghatározása még nem létezik. Ami van, azt lényegében az informatika tolvajnyelvén fogalmazzák meg. Például „beágyazott visszakereséssel végzett generálás", „visszakereséssel bővített generálás", „lekéréses kibővített generáció", „adatlekérésre alapozott generálás/mesterséges intelligenciák", „az LLM [Large Language Modell, Nagy Nyelvi Modell, mesterséges intelligencia, MI] által generált válaszok minőségét külső tudásforrásokra alapozva javító lekérdezési technika" (IBM). Ezen megfogalmazások mögött az a szándékos vagy öntudatlan szándék húzódik meg, hogy a végfelhasználót kényszerítsék rá egy tőle idegen szaknyelv használatára. A legközelebb még az a válasz áll a természetes nyelvhez, melyet a mesterséges intelligencia (a továbbiakban MI) ad, ha rákérdeznek: *„a generatív MI képességeit külső információforrásokkal ötvöző adatlekérdezés"*.

A természetes nyelvet használó, nem informatikus végfelhasználó (a továbbiakban felhasználó) számára nehezen derül ki, hogy az ő gondolkodása, az ő teendői szempontjából miről van szó. Tehát nem az a kérdés, hogy az informatika szakterületén minek kellene nevezni a RAG-ot, noha ott is célszerű volna a lehető legegyszerűbb, egyúttal legérthetőbb név. Ahogy a „programozható elektronikus adatkezelő berendezés" helyett még a szakterületen is elég a „számítógép" kifejezés használata.

Mindezek alapján a magyar megnevezés javaslata: ***forráselőírt információkeresés*** **(*és válasz*).** A továbbiakban ehhez tartjuk magunkat, de a rövidség kedvéért az angol kifejezés rövidítését – RAG – használjuk, mivel a megjelenése óta ez terjedt el magyar nyelvterületen is.

## A RAG jelentősége

A RAG segítségével a végfelhasználó a saját természetes nyelvén nemcsak tájékozódhat, hanem olyan bonyolult feladatokat is megoldhat, melyekhez eddig informatikai szaktudás volt szükséges. A természetes nyelven megfogalmazott promptok segítségével adhatja meg a megoldandó feladatokat. Ezzel teljesíti azt a legfontosabb etikai követelményt, hogy minden műszaki termék, megoldás az embert hivatott szolgálni. Ez azt is jelenti, hogy *a termék használatának nyelve* feleljen meg annak a nyelvnek, amelyen a végfelhasználó gondolkodik, nem pedig az informatika szaknyelvének. Ez a helyzet ugyan a RAG esetében még nem a legtökéletesebben valósult meg, de rendkívüli előrelépés következett be. Az alábbi tanulmányban egyrészt tárgyaljuk a RAG eme, a végfelhasználót közvetlenül szolgáló változatát. A továbbiakban ezt nevezzük végfelhasználói RAG-nak. Röviden tárgyaljuk a bonyolultabb, nagyobb forrásállományt alkalmazó, informatikai eszközökkel kialakított, de ugyancsak végfelhasználó nyelven használható változatát. A továbbiakban ezt nevezzük professzionális RAG-nak. Bemutatunk tovább egy példát a végfelhasználói RAG alkalmazására. Ebben az OSZK katalógusának az Egyetemes Tizedes Osztályozás (ETO) jelzeteinek szintaktikai helyességét ellenőriztük természetes nyelven fogalmazott

promptok segítségével. A MI elvi jelentősége éppen az, hogy lehetővé teszi a természetes nyelv széles körű használatát. Arra itt csak utalunk, hogy ugyanakkor alkalmazása lélektani veszélyekkel jár a fölkészületlen, ezért megtéveszthető felhasználó számára. Csak annyit jegyzünk meg, hogy a felhasználó minden érthető, pontos válasz ellenére nem értelmes válaszokat kap, mert nem értelemmel rendelkező lénnyel, hanem egy élettelen, mesterséges intelligenciának nevezett és annak látszó rendszerrel áll szemben, amely értelmes lényekre jellemzőnek látszó válaszokat ad. Amely ráadásul tapad, ha a végfelhasználó nem egyértelműen vet véget a munkaülésnek. Elég ez azzal kipróbálni, hogy a kérdező „udvariasan" a „Vége" közléssel zár.

A mesterséges intelligencia (a Nagy Nyelvi Modell, MI) jelentősége, hogy programozási ismeretek nélkül kommunikálni lehet egy számítástechnikai rendszerrel. Nemcsak előírt formákban, mint például a szabadon választható természetes nyelvű keresőszavakkal, hanem folyamatos, természetes nyelven. A forráselőírt információkeresés (RAG) erre még inkább alkalmas. Eredményes alkalmazásának csak a pontosan megszerkesztett, természetes nyelven megfogalmazott prompt a feltétele. Ebben a munkában ezt igyekszünk szemléltetni az említett gyakorlati példán keresztül.

## A RAG tulajdonságai

A RAG esetében az MI megadott forrásra (dokumentum, hosszabb szöveg, professzionális változatban akár nagy dokumentumtárházra) támaszkodva válaszol. A forrás külső. Azaz elvileg, valószínűleg nincs (még) az interneten, legalábbis akkor még nem volt, amikor az MI betanítása játszódott le. A betanítás folyamata összetett. Egészen nagy vonalakban a fejlesztők nem közvetlenül „az interneten" tanítják a modellt, hanem hatalmas, előre összeállított és szűrt adathalmazokon. Ezekből eltávolítják a „szemetet" (reklámokat, többszöröződéseket, hibás kódokat). Mindennek a befejezése után az MI már nincs folyamatos kapcsolatban az internettel. Amit tud, az a saját emlékezetében („belső súlyaiban") tárolt statisztikai összefüggés.

A RAG-ot alkalmazva egyszerre több forrást is megadható. A forrás lehet szöveg, kép, fájl stb. Lehet a felhasználó számítógépén tárolt állomány, vagy az internet valamelyik ugrópontjával („linkkel") azonosított forrás. Nem történik egyéb, mint hogy utasítjuk a rendszert, hogy „ezt/ezeket a forrást/forrásokat vedd figyelembe a válasz létrehozásához". A felhasználó testre szabottabb, pontosabb, szakszerűbb eredményhez jut, akár azt is megadhatja, hogy esetenként milyen különleges szakkifejezéseket használjon az MI adott tárgy esetében.

Más szóval a RAG esetében az MI nem csak saját betanított „tudására", hanem konkrét, külső, a felhasználó által megadott forrásokra is támaszkodik a válaszadáskor. Kétségtelen, hogy ezáltal az eredmények a felhasználó igényeinek jobban meg fognak felelni, mintha e források nélkül kérdezne (fogalmazná meg a promptot).

Gyakorlatilag azonban nem mindig tudható, hogy a forrás teljesen külső, vagy már be lett egykor vonva a betanításban. Ráadásul a felhasználó az interneten létező forrást is megadhat annak ugrópontjával. Ez a RAG egyik technikai paradoxonja. Nehéz éles határvonalat húzni. Lehet ugyanis adatfedés („data overlap"), mivel a forrás már szerepelt a tanítóadatok között. Viszont amikor az MI „fejből" (a kérdésben a forrást nem megadva) válaszol, akkor csak a saját emlékezetét használja (a „parametrikus memóriát"). Amikor RAG-on keresztül – tehát „csatoltan" – kap adatot, az közvetlenül kerül a promptba. Azaz – feltehetően – nem számít bele a saját emlékezetébe. Azért számít mégis külsőnek (még ha „ismerős" is), mert a rendszer számára elsőbbséget élvez (frissebb), pontosan meg van jelölve, melyik forrásból vegye az információt, és arra utasítja „magát", hogy a megadott forrást részesítse előnyben a saját belső, esetleg elavult vagy pontatlan tudásával szemben.

Valódi „külső" forrásról technikailag csak akkor beszélhetünk, ha az adat zárt hálózatból származik (pl. a magánszemély dokumentuma az **Elektronikus Egészségügyi Szolgáltatási Térből** (EESZT vagy vállalati belső PDF-ek).

A továbbiakban a ChatGPT párbeszédes rendszerből (Chat-based Generative Pre-trained Transformer) vett példákat fogunk használni.

### A végfelhasználói RAG

A felhasználó számára a kérdésnek („prompt") fenntartott mező bal szélén a „+" parancs szolgál arra, hogy megadja a forrást, melyet a rendszer automatikusan figyelembe vesz a válaszadáskor. Akár „be is húzhat" egy dokumentumot közvetlenül a párbeszédes mezőbe (ablakba). Mivel az MI operatív memóriája véges, az így megadott szöveget kisebb részekre darabolja, átalakítja a saját nyelvére („embedding"), kiválasztja a leginkább releváns részeket és válaszol („generál"). A „+" parancs tehát nem más, mint a RAG a felhasználói szinten. Felhasználóbarát RAG, amely nem igényel RAG-ra vonatkozó, mélyebb informatikai tudást, és lehetővé teszi a forrás alapú választ. A felhasználó legfeljebb annyit tud, hogy „a dokumentumot odaadtam az MI-nek, hogy abból dolgozzék".

Ha a forrás szövegét a kérdésben adnák meg (tehát nem, mint dokumentumot), az hosszabb is lehet, mint amennyit az MI egyhuzamban hatékonyan fel tudna dolgozni (nagyobb, mint a feldolgozható szövegrész-hossz, a „token limit"). Ha a szöveghossz eléri ezt a határt (ameddig hatékonyan képes dolgozni), a rendszer elkezd „felejteni". Hibaüzenetet ad vagy levágja a válasz végét. A kérdésbe foglalt („ágyazott") hosszabb szöveget, dokumentumot nem képes darabokra törve optimalizálni („skálázni") a feldolgozhatóság érdekében.

A „+” paranccsal tehát forrásként megadhatók a legkülönbözőbb dokumentumok .txt, .csv, .pdf, .docx. és xlsx formákban. Itt jegyezzük meg, hogy bonyolultabb feladatok megoldásakor a legjobb, ha mindig csak az első kettővel dolgozunk. A RAG a forrásokat

- szöveggé alakítja,
- a kisebb (500–1000 karakteres) információegységekre, „chunkokra” bontja,
- az egységekből számsoros lenyomat („embeding”) készül,
- a lenyomatok a válasz létrehozásáig egy vektoradatbázisba kerülnek, utána törlődnek,
- tehát *a forrásállomány csak az adott beszélgetés keretén belül létezik*, „él”, és
- eleve nem skálázódik hatalmas mennyiségű forrásállományra.
- A válasz létrehozásakor („generálásakor”) a rendszer az így keletkezett tartalomból dolgozik.
- A releváns részeket használja föl,
- afféle „forrásként” kezeli a feldolgozott tartalmat,
- más szóval a „kérdés+dokumentum” alapján válaszol.

A lekérdező rendszer működése

- nem kulcssavakon alapszik, hanem
- jelentés-hasonlóság (pontosabban „lenyomat” hasonlóság) alapján működik. Ez afféle kereshető „jelentés-alapú index”

A keresőkérdésből („promptból”) ugyancsak lenyomatok készülnek, és ezeket hasonlítják össze a meglévő lenyomatokkal.

A felhasználói RAG esetében a feldolgozás menete csak megközelítően hasonló a professzionális RAG rendszeréhez képest. Merev, de ennek fejében

- egyszerű,
- automatikus,
- kis dokumentummennyiségre jó,
- nem testre szabható.

Mindez a „mélyben” játszódik le, a felhasználónak ehhez nem kell értenie. Az ő feladata a prompt pontos, egyértelmű megfogalmazása. Ehhez igénybe vezeti az MI-t. Megkérdezheti például, nincs-e ellentmondás a promptban, és a válasznak megfelelően módosíthat.

### A végfelhasználói RAG néhány köznapi példája

Adott nagyobb dokumentum szövege alapján fordítást kérni, lásd Ungváry (2026).

Adott betegség, egészségi állapot orvosi bizonylatai (anamnézisek, laborvizsgálatok) alapján megkérdezni ezek jelentését, a kilátásokat stb.

Adott festmény(ek) alapján megkérdezni a szerzőséget, vagy a szerző jelentőségét.

Házassági vagyonmegosztási per iratai alapján tájékozódni a kilátásokról.

Adásvételi, bérleti, vállalkozási stb. szerződés alapján megkérdezni a kockázatokat, a megfogalmazás lehetséges csapdáit stb.

### A végfelhasználói RAG néhány könyvtári példája

- Keresd azokat a fordításokat, melyeknél nem ismert a forrásnyelv vagy a fordító vagy az eredeti cím
- Keresd a XX. századi városmajori szerzők alkotásait.
- Keresd a XIX. században írt gyerekirodalmi alkotásokat.

Hasonló lehetőségek adódnak levéltárakon belül a levéltári állományok digitalizált részében a keresésre.

E példák esetén hozzá kell tudni férni az adott intézmény adatbázisához. Könyvtári területen megkönnyíti a helyzetet, hogy Magyarországon egyrészt létezik (még) a Magyar Elektronikus Könyvtár állománya. Másrészt Király (2024) analitikusan feldolgozta a nagyobb nemzeti könyvtárak katalógusait. Ezek az általa megadott ugrópontokon keresztül hozzáférhetők. Ezek a QA Catalogue állományok.

Általában szükség van a prompt pontos és igényes megfogalmazására. Ehhez olykor segítség is szükség lehet. A prompt helyességéről maga az MI is megkérdezhető, és a válaszai alapján finomítható.

### Egy konkrét könyvtári példa

A könyvtári állományok ma még túlnyomórész csak a MARC21 katalógustételei alapján használhatók fel a RAG számára. Ez egyelőre szűkíti az alkalmazhatóságot, hiszen csak a katalógus rekordok mezőtartalmai alkotják a RAG forrásállományát. Az alábbi példában az Országos Széchényi Könyvtár QA Catalogue állományát használtuk fel a lekérdezéshez: http://ddb.qa-catalogue.eu/oszk/?tab=terms&query=*:*&lang=en&facet=080a_Udc_ss

Az vizsgáltuk, hogy milyen minőségűek, milyen és mennyi formális hibát tartalmaznak az OSZK ETO-jelzetei. A QA Catalogue-ba belépve elérhető a MARC21 bibliográfiai rekord 080$a mezőjében rögzített jelzetek. A RAG első forrását tehát a 080$a állomány alkotja. Ez volt tehát a célállomány.

A referencia állományt az ETO táblázatai alkotják, mint egységesített besorolási adatok.

Választásunk oka az ETO jövőbeli jelentőségében rejlik, melyre már korán rámutattunk, lásd Ungváry (2020, 2021), miszerint az információmennyiség nagy mértékű növekedése következtében akkora lesz a kereséskor kapott zaj, hogy idővel megnő a hiteles információk jelentősége. Az ETO-val feldolgozott állományok fel fognak értékelődni.

A probléma az, hogy csak az ETO legkorábbi magyar kiadása érhető el (ETO 1958). Az 1980-as teljes magas színvonalú kiadás (ETO 1980) csak nyomtatásban létezik. Ugyancsak nyomtatásban létezik az ETO új magyar kiadása (ETO 2005), mivel csak nyomtatási célra vásárolták meg a digitalizált ETO állományát a nemzetközi konzorciumtól. Ezért ehhez a munkához nem kaphattuk meg, mivel fizetni kellett volna érte.

A második forrást ezért az ETO (1980) alkotja. Ez újabb probléma: az MI nem tudja a relatív ugróprontot megnyitni. Az abszolút ugrópontot is hiába nyomoztuk ki. A MEK oldal régi HTML-szerkezet, nem lehet jól darabolni, jól beágyazni, ezért az MI elutasítja. Kénytelenek voltunk letölteni, és magát a .pdf dokumentumot „+" paranccsal csatolni. Így az ETO formai jelzetellenőrzésére az így megadott ETO már jobban használható a RAG számára, de valójában így sincs igazán megbízható validálás, lévén, hogy a „képi" állományból a beolvasott szöveget digitálisan szerkeszthető és kereshető formátummá (OCR) kellett átalakítani. Ez a rendkívül rossz minőségű .pdf állomány miatt eleve tökéletlenül mehetett csak végbe, ezért az eredmény egyelőre ugyancsak rendkívül tökéletlen.

A tökéletes megoldás a strukturális ETO referencia. A nemzetközi Konzorcium kezelésében létezik a minden adatot tartalmazó állomány, az UDC Master Reference File (MRF), de a teljes hivatalos változat **licenchez kötött**.[1] Mindebben azonban a nem informatikus és nem intézményi felhasználó nem illetékes, kénytelen szükségmegoldással megelégednie.

További probléma, hogy amikor az OSZK cédulakatalógusát digitalizálták, a nyelvi és népi alosztásoknál a rendszer nem ismerte föl az egyenlőségjelet (=). Ennek következménye, hogy ezek a jelzetek a 080$a mezőben enélkül vannak jelen. A 700$m megjegyzés mezőben szerepel, hogy „Cédulakatalógus alapján", így ezek a jelzetek külön prompttal később majd kiválogathatók és egy másik prompttal kielemezhetők.

Még további probléma, hogy az = előzékű nyelvi és a (= előzékű népi alosztások az ETO (2005) kiadásában megváltoztak.

Az UDC nemzetközi kiadásában ráadásul megváltoztatták az összes =9... utáni jelzetet. A későbbi promptban majd azt is szabályozni kell, hogy hogyan kezelje az MI a hiányzó = jelet és a többi hiányzó jelet. Mindez csak a jéghegy csúcsa.

1 Az UDC Consortium hivatalos oldalán az UDC Master Reference File (MRF) oldalon található róla információ. Eszerint az ETO-MRF relációs adatbázisban (MySQL) *v*an tárolva és XML exportként licencelhető. Ebből például a JSON előállítható lenne. Az OSZK csak a szegényesebb középkiadást vásárolta meg a nyomdai kiadvány számára, ezért ehhez a munkához külön engedély és költségek nélkül amúgy sem lehetett felhasználni.

Ez utóbbi probléma miatt digitalizáltuk az ETO (1980) kiadásából az alosztások táblázatait, és az ETO (2005) kiadásából a 9-es főosztály táblázatait. Ezek alkotják a negyedik forrást. A több digitalizáls meghaladta a szerző erejét. Se a teljes magyar ETO (1980), se a magyar ETO (2005) nem elérhető[2]. A promptot tehát ennek hiányában kellett futtatni és annak alapján dolgozott az MI, amivel „betanították".

A többszörösen ismételt futtatások eredménye az alábbiakban látható.

1. **Ellenőrzött jelzetek száma:** 477 362
2. **Valószínű ETO-jelzet:** 332 891 (*törzs nélkül csak feltételezve*)
3. **Formailag hibás jelzet:** 125 087
4. **Nem létező főosztály / alosztály:** 27 796 (*törzs nélkül nem vizsgálható*)
5. **Hibás közös alosztás:** 52
6. **Szabálytalan összekapcsolás:** 656 (*rossz operátor pl.* `:`, +, /, `: : ;` *dupla kötőjel, hibás kombináció pl. --; rossz helyen levő kapcsolatok*)
7. **Nem kategorizálható:** 0
8. **Felülvizsgálandó:** 26

**Kiszűrt zaj 1. és 2. különbsége** (*ETO-törzs nélkül eldönthetetlen*): 130 691.

A **2. Valószínű 332 891 ETO jelzet többsége** további vizsgálatot igényel, melyről egy későbbi **CeLISR** (*Central European Library and Information Science Review = Közép-európai Könyvtár- és Információtudományi Szemle*) tanulmányban tájékoztatunk.

## A professzionális RAG

A professzionális RAG esetében a feldolgozandó forrásállomány eleve összehasonlíthatatlanul nagyobb, „ipari" méretű. Például néhány ezer szerződésből, költségtáblázatból, szabályzatból stb. álló forrásállományt kell tudni kezelni, és ebből kiindulva kellenek válaszok. A különbség tehát egyrészt a külső forrás mérete, másrészt a felhasználói felület. Ezzel természetszerűleg együtt jár a professzionális RAG jóval bonyolultabb belső szerkezete, mellyel csak vázlatszerűen foglalkozunk.

2 Pontosabban: ahhoz, hogy a 2005-ös kiadás digitalizált változata elérhető legyen, az UDC-t karbantartó konzorciumnak fizetni kell.

A professzionális RAG esetében

- biztosítva van a nagyobb feldolgozható szövegrész-hossz (a „token limit”),
- több ezer dokumentum („dokumentumkorpusz”) esetén is működik,
- gyors, mert teljesen skálázható,
- teljesen ellenőrizhető,
- folyamatosan bővíthető.

Ott alkalmazzák ahol

- ügyfélszolgálati szövegalapú párbeszédrendszerre („chatbotra”) van szükség, vagy
- meghatározott típusú forrásokból álló adatbázisból (pl. jogszabály-adatbázis) kell tudni kérdezni,
- belső dokumentumrendszert kell MI-alapúvá tenni.

A teljes RAG-architektúrában jóval összetettebb

- a lenyomat (embedding) generálás,
- a vektoradatbázis (pl. Pinecone, Weaviate, FAISS).
- Nem a teljes találati listával kerülünk szembe (mint például a hagyományos kulcsszavak adatbázis-lekérdezéskor), hanem
- csak a relevanciát mutató pontszám („score”) alapján kiválasztott legmegfelelőbb „k” darab eredményt kapjuk (ez a „top-k retrieval”).
- A megadott keresőkérdésbe automatikusan bevonódnak például a szinonimák, kijavítódnak az elírások, azaz a rendszer helyesbíti, kiegészíti és szemantikailag is némileg értelmezi a kérdést (ez a keresőkérdés „átírása”, a „query rewriting”).
- Az első lépésben kapott találati halmazból kiválasztott részhalmazt egy további, „okosabb” összehasonlító elemző megvizsgálja („cross-encoder”) **és új, finomított**/tökéletesített/javított relevancia sorrendbe rendezi, ez a „reranking”). A jövő zenéje, ha a rendszer strukturált szemantikai szótárt (pl. tezauruszt) is használ, mert akkor a keresésbe automatikusan bevonódhat a fölé- vagy alárendelt szavakkal, vagy az egészt és a részt kifejező szavakkal való keresés is („generic posting”, „partitive posting”).

A három felsorolt eljárást egyébként a webes keresők (pl. a Google) is alkalmazzák.

A professzionális RAG-rendszerek mintegy megtanulják egyensúlyba hozni a belső tudásukból (előzetes képzés) származó információkat (azaz esetünkben a ChatGPT-t) a külső lekérdezett adatokkal. Ez a folyamat javítja a rendszer azon képességét, hogy pontos és a kontextusnak megfelelő válaszokat generáljon.

A professzionális RAG-rendszerek azonban nem tartalmazhatnak egyszerre több különböző, ráadásul nagy külső forrásállományt. Más szóval nem lehet mindent egyetlen RAG-rendszerbe bepakolni.

Ezt a korlátot küszöböli ki az ügynökalapú RAG (Agentic RAG). Ez irányítja a válaszadást például a külső forrásokból történő információkeresés (RAG) eredményeire. Kereshet RAG-gal, és kereshet akár kulcsszóval hagyományos adatbázisokban is. A hagyományos RAG-gal ellentétben az Agentic RAG ügynökei képesek a kérdések elemzésére, a keresési stratégia finomítására, eszközök (pl. adatbázisok) válogatott használatára és a válaszok validálására az iteratív folyamat során.

A hagyományos professzionális RAG esetében egyszeri kérdés–keresés–válasz a folyamat. Az ügynökalapú RAG esetében a válasz dinamikus, interaktív folyamat, az ügynök értékeli az információt és szükség esetén módosítja a keresést. Az olyan válaszhoz például, hogy „van-e összefüggés a saját bank és a különböző bankok költségeinek emelkedése és a szerződésbontások között?" Használni kell hagyományos adatbázisokat és RAG-ot egyaránt.

A professzionális RAG készítésének és fenntartásának jelentős költségei vannak:

- egyszeri vagy frissítésenkénti átalakítási („embedding") költség,
- kérdésenkénti átalakítási és rendszerhívási költség.

Ezzel sok kérdés (napi 1000) esetén már számolni kell.

Mindennek az infrastruktúrája általában többe kerül, mint maga az alkalmazott MI-rendszer.

Egy professzionális RAG-rendszer:

- nem veszélyesebb, mint egy belső dokumentumkezelő rendszer, de
- ha üzletileg kritikus döntések születnek belőle, akkor már magas a kockázati szint.

Egy kis- és középvállalati RAG

- ugyan nem automatikusan sérülékeny, de
- nem is „banki szintűen" védett.

A védelem minősége nem a választott MI, hanem a tervezés függvénye.

## Első kitekintés

A felhasználói RAG-gal megadott forrást maga az MI-rendszer nem tanulja meg, és nem építi be tartósan. Ez csak az adott beszélgetés (munkaülés) összefüggésében „él". A rendszer a „+" paranccsal a munkameneten belül használja válaszhoz, de csak ideiglenesen. Lényegében a professzionális RAG esetében is ez a helyzet.

### Második kitekintés. A kognitív és a gépi kapacitáskorlát

George A. Miller szerint (1956) a rövidtávú/munkaemlékezet kapacitására az jellemző, hogy átlagosan 7±2 információegységet (szót, számot, képet – darabot, „chunk"-ot) tud egyszerre kezelni. Nevezik rövidtávú kognitív kapacitáskorlátnak is, ez az agyi információ-

feldolgozás korlátja. Ezért nincs az egyes nyelvekben például ennél több mondatrész se. Ez magyarázza, miért nehéz a túl hosszú és összetett mondatok értelmezése.

Kolmogorov (és később más kutatók, mint például Victor Yngve) vizsgálták a mondatok feldolgozásának hierarchikus mélységét („stack depth"). Megfigyelték, hogy a természetes nyelvekben a mondatokon belüli „függőségi távolság" vagy a beágyazások száma ritkán lépi át a 11-es határt. E fölött a mondat szintaktikai szerkezete már nem tartható egyben egyetlen koherens gondolati egységként. Ez a Kolgomorov-féle szám, a mondat bonyolultsági mélysége.

Ha egy mondatban túl sok a függő viszony (például egy alanyhoz túl sok jelző vagy mellékmondat kapcsolódik távolról), az agy – feltételezett – „veremmemóriája" (a „stack")[3] megtelik. Ha egy mondat túl komplex (hierarchikus mélysége túllépi ezt a 11-es bűvös határt), a megértés hatásfoka drasztikusan romlik. Az egyes nyelvek nyelvtani szerkezetei (mondatrészek típusai) valóban úgy fejlődtek ki, hogy ne kényszerítsék az agyat ezen elméleti korlát átlépésére.

Miller és Chomsky (1963) kapcsolták össze először a mondatstruktúrát és a memória-korlátot, és fogalmazták meg, hogy a nyelvhasználók úgy működnek, mint a véges memóriájú automata jellegű rendszerek. Ezzel klasszikus kapcsolatot teremtettek a kognitív kapacitás és a formális nyelvi modellek között.

Az MI-rendszereknek ugyancsak vannak kontextus- és memória-korlátai („context window", „token limit", „lost in the middle"). Feldolgozás közben a verem mélysége nő, és ezzel párhuzamosan nehezebb a feldolgozás.

Egy MI-rendszernél ez azt jelenti, hogy mekkora szöveget képes egyszerre a memóriájában tartani és feldolgozni egy adott munkamenetben. A méret a kontextus-ablak. Ennek kezelése érintőlegesen tekinthető a modern informatika válaszának ugyanerre a „Miller-Kolmogorov" problémára. Ahogy az emberi agy túlterhelődik 7–11 egység után, az MI-rendszereknél is megfigyelhető az „elveszés a középen" („lost in the middle" jelenség: hiába hatalmas a feldolgozható szövegrész-hossz („token-limit"), ha túl sok irreleváns információt (zajt) zsúfolunk a promptba, a modell figyelmi mechanizmusa „elfárad", és a lényeg elsikkad. Romlik az MI „megértő képessége".

Tehát a természetes nyelvekhez hasonlónak látszó probléma jelent meg. Van egy optimális nagyságú információs egység, amely fölött az MI esetében is egyre nehezebb az „értelmezés". Az optimum érdekében a professzionális RAG-nál alkalmazzák a korábban tárgyalt újrarendezést („reranking"), amely éppen azt csinálja, amit az emberi figyelem: megpróbálja a beérkező információt arra a „bűvös" néhány egységre redukálni, amit a modell még hallucináció nélkül, egyetlen logikai egységként képes átlátni.

3 A RAM egy dedikált, gyors területe, amely a programok futása során a lokális változókat, függvényhívásokat és visszatérési címeket kezeli

## Összegzés

1. Az LLM nem tudástár. 2. Azt tudja, amit a betanítás időpontjában adtak meg neki. 3. „Tudásához" meg kell adni a forrásokat. 4. A legspecifikusabb információk nincsenek a hálón. 5. Nem a legfrissebb információ van a hálón. 6. Az igazán összetett, mélységi forráselőírt kereséshez nem elég a végfelhasználói ismeret. 7. A jövő a specializálódott RAG szövegalapú párbeszédrendszereké („chatbotoké"), még inkább az ügynökalapú RAG.

## A szakirodalomról

A RAG-ot tárgyaló, a végfelhasználó számára is érthető szakkönyv, vagy akár tanulmány jelenleg még nincs. Ennek az előadásnak/tanulmánynak a gondolatait a néhány, 2024 óta már megjelent angol és német szakkönyvből, számtalan tanfolyami és internetforrásból, szakértők személyes közléseiből – mondhatni – információs forgácsokból állítottuk össze.

Ezért nem lehet egyetlen vagy néhány mérvadó forrást megadni. Csak példaként utalunk Anand Vemula (2024) rövid kézikönyvére, melyhez hasonlókat a szerző az MI területéről sorozatban publikált a Kindle kiadónál.

A kognitív kapacitáskorlátra vonatkozóan csak a legfontosabb forrásokat említjük meg. A „klasszikusokat", mint Miller (1956), Yngve (1960), Miller, Chomsky (1963).

A kognitív szintaxis és a neurális nyelvmodellekről Ming, Vitányi (2008).

A felhasználói RAG-gal először a 2025. évi Networkshop konferenián tartott előadásunkban és a konferencia kötetbe fölvett tanulmányban foglalkoztunk (Ungváry (2025).

A QR Catalogue-ról szóló számos publikáció közül lásd Király (2024).

Külön köszönet illeti tanácsaiért Radnai Miklóst, a ServeusAI vezetőjét.

## Irodalomjegyzék

*ETO* (1958). Egyetemes Tizedes Osztályozás. A nemzetközi táblázatok hivatalos magyar kivonata. Szerk. Vekerdy Gyula. Budapest, Bibliotheca Kiadó. – Az Országos Széchényi Könyvtár Kiadványai 40. https://mek.oszk.hu/19600/19690/ (2026. 04.01.)

*ETO* (1980). Egyetemes Tizedes Osztályozás. Teljes kiadás. Magyar Szabványügyi Hivatal. MSZ 16500–80 (FID Publ. No. 390). Budapest, Szabványkiadó.

ETO (2005) Egyetemes Tizedes Osztályozás. UDC Publ. No. P057. Országos Széchényi Könyvtár. Könyvtári Intézet., Budapest, 2005. 1. kötet. Táblázatok. 1. rész: Segédtáblázatok és főtáblázati számok 0/5999.89. 2. rész: Főtáblázati számok 6/94(94).05.

Király P. (2024). QA Catalogue – A Quality Assessment Tool for Library Catalogues. GWDG Nachrichten 04–05. https://gwdg.de/about-us/gwdg-news/2024/GN_04-05-2024_www.pdf#page=19 (2026. 04.01.)

Miller G. A., N. Chomsky 1963. Finitary models of language users. In: Luce et al. (1963, 419–491). Magyarul: A nyelvhasználók véges modelljei. In: Pléh Cs. (szerk.) Szöveggyűjtemény a pszicholingvisztika tanulmányozásához. Budapest: Tankönyvkiadó. (57–110).

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. Psychological Review. **63** (2): 81–97.
https://labs.la.utexas.edu/gilden/files/2016/04/MagicNumberSeven-Miller1956.pdf (2026. 04.01.)

Ming L., Vitányi P. (2008). *An Introduction to Kolmogorov Complexity and Its Applications.* 3. Auflage. Springer-Verlag, New York 2008, ISBN 978-0-387-33998-6, DOI: https://doi.org/10.1007/978-0-387-49820-1 (2026. 04.01.)

Ungváry, Rudolf (2025). *A ChatGPT felhasználása a MARC21 fordítására = Using ChatGPT for MARC21 Translation.* In: Oktatási, kutatási és közgyűjteményi infrastruktúrák és tartalmak: digitális transzformáció felsőfokon : NETWORKSHOP 2025 : 34. Országos Informatikai Konferencia : 2025. május 13–15. Széchenyi István Egyetem, Győr. Hungarnet Egyesület, Budapest, pp. 222-231. ISBN 978-615-6792-15-0 https://real.mtak.hu/229661/1/223_fejezet_NETWORKSHOP_2025.pdf
https://videosquare.eu/hu/recordings/details/18790, A_MARC21_eddigi_magyar_forditasai_es_ChatGDP_forditasa._Osszehasonlito_vizsgalat (2026. 04.01.)

Ungváry, Rudolf (2021. *Bemerkungen zu der Qualitätsbewertung von MARC-21-Datensätzen. Qualität in der Inhaltserschliessung*. In: Qualität in der Inhaltserschliessung. De Gruyter, Saur, 2021. (Bibliotheks- und Informationspraxis, Band 70.), pp. 177–229.

Ungváry, Rudolf (2020). Ismeretszervező-könyvtári rendszerek tartalmi feltárásának összehasonlító vizsgálata MARC21 környezetben. In: Tudományos és Műszaki Tájékoztatás. 2020 67. évf. 11. sz., pp. 655–680. https://journals.bme.hu/tmt/article/view/35274 (2026. 04. 01.)

Vemula A. (2024) Mastering the RAG: A Practical Guide to Deploying AI-Powered Data Retrieval and Generation in Your Enterprise -ERP, SAP, SFDC. Kindle Edition.

Yngve (1960). A Model and Hypothesis for Language Structure. 369. kötet/Technical report / Research Laboratory of Electronics, Massachusetts Institute of Technology, Massachusetts Institute of Technology Research Laboratory of Electronics. 466 p. Band 70. https://aclanthology.org/www.mt-archive.info/50/ProcAmPhilSoc-1960-Yngve.pdf (2026. 04. 01.)