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.2026.16

MESTERSÉGES INTELLIGENCIA ASSZISZTENS FELHŐALAPÚ RAG ARCHITEKTÚRÁBAN

ARTIFICIAL INTELLIGENCE ASSISTANT IN A CLOUD-BASED RAG ARCHITECTURE

Kelemen László Ádám
Programozó, Azure Cloud & AI Engineer
Szegedi Tudományegyetem, Informatikai Fejlesztési Igazgatóság
kelemen.laszlo.adam@szte.hu

Absztrakt

A nagy nyelvi modellek fejlődése lehetővé tette olyan intelligens asszisztens rendszerek létrehozását, amelyek természetes nyelven támogatják a felhasználókat az információkeresési és ügyintézési feladatok során. A kizárólag generatív modellekre épülő megoldások ugyanakkor intézményi környezetben korlátozott megbízhatóságúak lehetnek, mivel pontatlan vagy nem naprakész válaszokat is előállíthatnak. A tanulmány egy felhőalapú *Retrieval-Augmented Generation* (RAG) architektúrára épülő mesterséges intelligencia asszisztens keretrendszert mutat be, amelyet a Szegedi Tudományegyetem számára fejlesztettünk. A rendszer Azure felhőszolgáltatásokra épül, és hibrid keresést, vektoralapú dokumentum-visszakeresést, szemantikus újrarangsorolást, valamint többfázisú dokumentumfeldolgozást alkalmaz. A dokumentumfeldolgozás során jelenleg két chunking stratégia támogatott: a standard *paragraph-first chunking* és a fejezetszintű struktúrát figyelembe vevő *hierarchical chunking*. A rendszer enterprise retrieval optimalizációkat, jogosultságkezelést és biztonsági mechanizmusokat is tartalmaz. Az eredmények alapján az architektúra alkalmas nagyméretű intézményi dokumentumállományok feldolgozására és kontextusérzékeny válaszok generálására, miközben csökkenti a hallucinációs kockázatot.

Kulcsszavak: mesterséges intelligencia, Retrieval-Augmented Generation, RAG, hibrid keresés, vektorkeresés, dokumentum-visszakeresés, felhőalapú architektúra, Microsoft Azure

Abstract

The advancement of large language models has enabled the development of intelligent assistant systems that support users in natural language information retrieval and administrative tasks. However, solutions relying exclusively on generative models may have lim-

ited reliability in institutional environments, as they can produce inaccurate or outdated responses. This paper presents a cloud-based *Retrieval-Augmented Generation* (RAG) artificial intelligence assistant framework developed for the University of Szeged. The system is built on Microsoft Azure cloud services and employs hybrid search, vector-based document retrieval, semantic reranking, and a multi-stage document processing pipeline. During document processing, the system currently supports two chunking strategies: standard *paragraph-first chunking* and *hierarchical chunking*, which preserves chapter-level semantic structure. The architecture also includes enterprise retrieval optimizations, access control, and security mechanisms. The results indicate that the system is suitable for processing large institutional document collections and generating context-aware responses while reducing hallucination risk.

Keywords: artificial intelligence, Retrieval-Augmented Generation, RAG, hybrid search, vector retrieval, document retrieval, cloud architecture, Microsoft Azure

1. Bevezetés

A mesterségesintelligencia-alapú asszisztensek fejlődése szorosan kapcsolódik a nagy nyelvi modellek és a Transformer architektúra elterjedéséhez (Vaswani, és mtsai., 2017). Ezek a rendszerek természetes nyelven képesek kommunikálni, kérdésekre válaszolni, valamint információkeresési és ügyintézési folyamatokat támogatni. Intézményi környezetben azonban a kizárólag generatív modellekre épülő megoldások nem minden esetben elegendőek, mivel válaszaik nem feltétlenül ellenőrizhetőek, naprakészek vagy forrásokkal alátámasztottak.

A problémára megoldást jelenthetnek a *Retrieval-Augmented Generation* (RAG) architektúrák, amelyek a generatív modelleket dokumentumalapú információ-visszakereséssel egészítik ki (Lewis, és mtsai., 2020). A RAG-rendszerek a felhasználói kérdéshez releváns dokumentumrészleteket keresnek vissza, majd ezeket kontextusként adják át a nyelvi modellnek. Ez csökkentheti a hallucinációk előfordulását, és lehetővé teszi naprakész intézményi tudásbázisok használatát.

Jelen tanulmány a Szegedi Tudományegyetem számára fejlesztett, felhőalapú mesterséges intelligencia asszisztens keretrendszert mutatja be. A rendszer Microsoft Azure szolgáltatásokra épül, és komplex RAG pipeline-t valósít meg dokumentumfeldolgozással, vektoralapú indexeléssel, hibrid kereséssel, szemantikus újrarangsorolással, valamint enterprise szintű jogosultságkezeléssel. A fejlesztés célja olyan architektúra kialakítása volt, amely nagyméretű intézményi dokumentumállományok esetén is skálázható, biztonságos és alacsony hallucinációs kockázattal működik.

A tanulmány külön hangsúlyt helyez a dokumentumfeldolgozási stratégiákra. A rendszer két chunking (szeletelés) módszert támogat: a *paragraph-first chunking* megközelítést,

amely a paragrafusokat tekinti elsődleges szemantikai egységnek, valamint a *hierarchical chunking* stratégiát, amely a dokumentum fejezetszintű szerkezetét is figyelembe veszi. A további fejezetek bemutatják a rendszer architektúráját, a dokumentumfeldolgozási láncot, a *retrieval pipeline* (visszakeresési folyamat) működését, valamint a biztonsági és enterprise komponenseket.

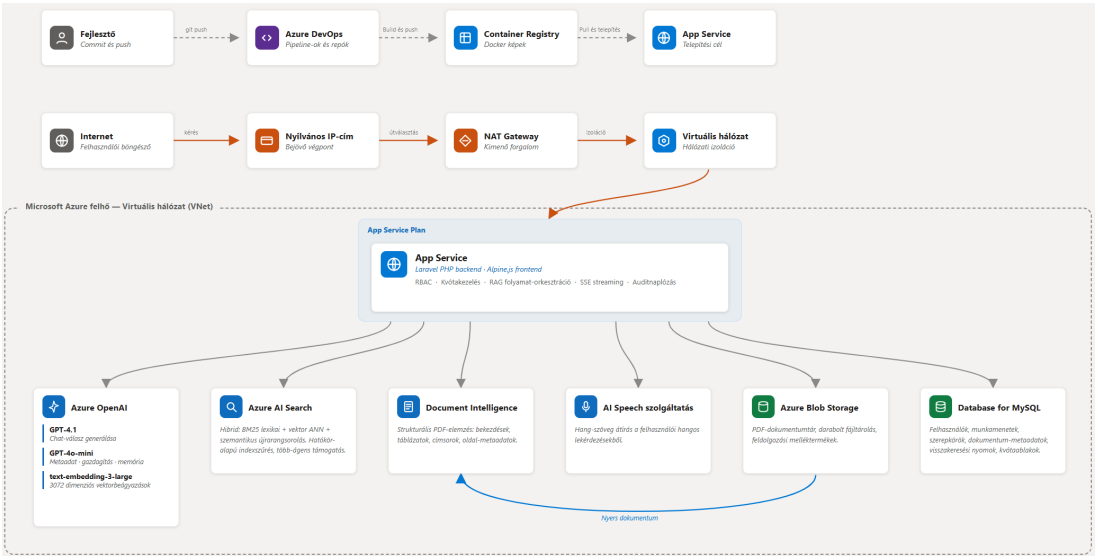
2. A rendszer architektúrája

A fejlesztett AI-asszisztens keretrendszer egy felhőalapú, többkomponensű RAG-architektúrára épül. Elsődleges célja intézményi dokumentumállományok természetes nyelvű lekérdezése és forrásalapú válaszok generálása. A rendszer kialakításánál kiemelt szempont volt a skálázhatóság, a modularitás, a dokumentum-hozzáférések kontrollja és az enterprise környezetben elvárt biztonság.

Az infrastruktúra Microsoft Azure szolgáltatásokat használ. A válaszgenerálást Azure OpenAI biztosítja GPT-alapú modellekkel, a dokumentum-visszakeresést Azure Search végzi, míg a PDF dokumentumok strukturális feldolgozását Azure Document Intelligence támogatja. A dokumentumok és kapcsolódó feldolgozási állományok Azure Blob Storage-ban tárolódnak. Az alkalmazás backend rétege Laravel alapú PHP-környezetben, a felhasználói felület pedig Alpine.js és Tailwind CSS technológiákkal készült. A rendszer deployment folyamata konténerizált üzemeltetési modellt alkalmaz, amely Azure Web App környezetben biztosítja a teszt- és éles infrastruktúra skálázható működését. A rendszer strukturált adatainak perzisztens tárolását Azure Database for MySQL szolgáltatás biztosítja.

A rendszer működésének központi eleme a dokumentumfeldolgozási és visszakeresési folyamat. A feltöltött dokumentumok metaadat-generáláson és strukturális elemzésen mennek keresztül, majd szeletelési stratégiák alapján dokumentumszeletekre bomlanak. A szeletek *embedding* (vektoros) reprezentációi a text-embedding-3-large modell segítségével készülnek (OpenAI, 2024), majd Azure Search indexbe kerülnek. A felhasználói kérdések feldolgozásakor a rendszer hibrid keresést alkalmaz, amely a BM25-alapú lexikális keresést, a vektoralapú szemantikus keresést és a szemantikus újrarangsorolást kombinálja (Robertson & Zaragoza, 2009).

Az architektúra több intelligens ágens párhuzamos működését támogatja. Az ágensek eltérő dokumentumkészletekhez, keresési scope-okhoz (hatókör) és jogosultsági szintekhez rendelhetők, így különböző intézményi területek elkülönítetten kezelhetők. A szerepalapú hozzáférés-vezérlés biztosítja, hogy a visszakeresési folyamat csak a felhasználó számára engedélyezett dokumentumokon fusson.



1. ábra: Az MI-alapú RAG-asszisztens rendszerarchitektúrája Microsoft Azure környezetben.

3. Dokumentumfeldolgozás és chunking (szeletelési) stratégiák

A RAG-rendszerek hatékonyságát jelentősen befolyásolja az indexelt dokumentumszeletek minősége. A túl nagy szeletek ronthatják a visszakeresés pontosságát, a túl kicsik pedig kontextusvesztést okozhatnak. Ezért a rendszerben a dokumentumfeldolgozás nem pusztán technikai előkészítő lépés, hanem a visszakeresés minőségét meghatározó komponens.

A feldolgozási folyamat a PDF-állományok feltöltésével indul. A dokumentumok Azure Blob Storage-ba kerülnek, majd automatikus metaadat-generálás történik, amely a dokumentum címét és rövid leírását állítja elő. A strukturális elemzést Azure Document Intelligence végzi, amely paragrafusokat, táblázatokat és egyéb logikai egységeket azonosít. Ezekhez oldalszám, szerkezeti szerep és egyéb metaadatok társulhatnak.

A rendszer jelenleg két szöveges szeletelési stratégiát támogat. A standard *paragraph-first chunking* a paragrafusokat önálló szemantikai egységként kezeli. A módszer nem alkalmaz oldalhatár-alapú szeletelést, így az összetartozó tartalom akkor is egy chunkban maradhat, ha több oldalra tördelődik. Ez különösen hasznos szabályzatok, tanulmányi dokumentumok és ügyintézési leírások esetén, ahol a fejezetcím és a hozzá tartozó tartalom gyakran eltérő oldalon jelenik meg.

A paragraph-first stratégia puha és kemény karakterkorlátokat használ. A puha korlát a preferált szeletméretet jelöli, míg a kemény korlát a rendkívül hosszú paragrafusok kezelését biztosítja. A rendszer átfedéses szeletelést is alkalmaz, amelyben a szomszédos chunkok bizonyos közös paragrafusokat tartalmazhatnak, ezzel javítva a kontextus folytonosságát.

A *hierarchical chunking* stratégia a dokumentum fejezetszintű szerkezetét is figyelembe veszi. A módszer az azonos logikai egységhez tartozó paragrafusokat közösen kezeli, és a

chunkokhoz fejezethierarchiát kapcsolhat. Ez különösen előnyös nagyméretű, strukturált dokumentumok esetén, ahol a fejezetszintű kontextus megőrzése javíthatja a visszakeresés relevanciáját.

A táblázatos tartalmak feldolgozására külön mechanizmus szolgál. A táblázatok soralapú szeletekre bonthatók, miközben a fejlécinformációk ismétlése biztosítja, hogy az önállóan visszakeresett szeletek értelmezhetők maradjanak. Ez tantervi követelményeket, kreditstruktúrákat vagy adminisztratív adatokat tartalmazó dokumentumok esetén kiemelten fontos.

A feldolgozott dokumentumszeletek vektoros formában kerülnek indexelésre (OpenAI, 2024). Az vektoros generálás kötegelt módon történik, és gyorsítótárazás támogatja, amely csökkenti az ismételt API-hívások számát. A többféle szeletelési stratégia célja, hogy eltérő dokumentumszerkezetek esetén is szemantikailag koherens és hatékonyan visszakeresheto egységek jöjjenek létre.

3.1 A két szeletelési stratégia összehasonlítása

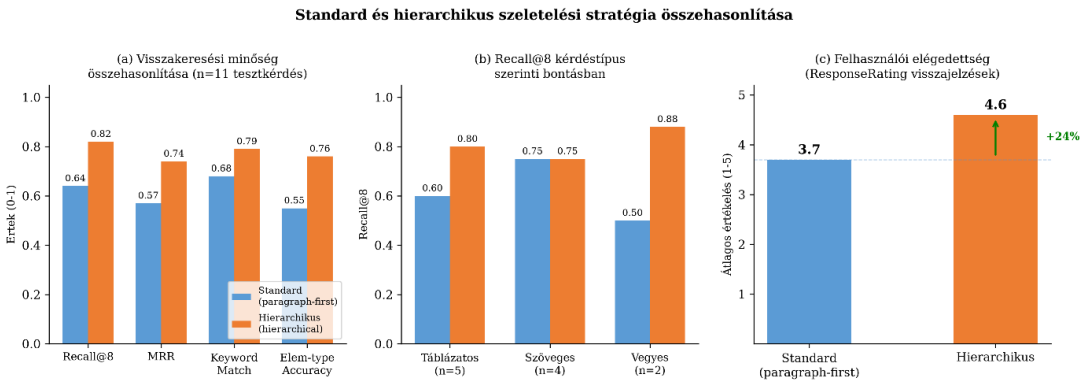
A 2. ábra a standard bekezdésalapú (paragraph-first) és a hierarchikus (hierarchical) szeletelési stratégia összehasonlítását mutatja be három dimenzióban, 11 reprezentatív tesztkérdés alapján.

Az a) diagram négy visszakeresési mutatót szemléltet 0–1 skálán: a *Recall@8* a helyes dokumentum top 8 találat közötti előfordulási arányát; az *MRR* (Mean Reciprocal Rank) a helyes találat átlagos ranghelyének reciprokát (1,0 = mindig első); a *Keyword Match* az elvárt kulcsszavak visszakeresési lefedettségét; az *Element Type Accuracy* a visszaadott tartalom típusának (tábla/szöveg) helyességét méri. A hierarchikus stratégia mindegyik mutatón felülmúlta a standard megközelítést.

A b) diagram a *Recall@8* értékét bontja le kérdéstípusonként: táblázatos lekérdezéseknél a hierarchikus chunking 0,60-ról 0,80-ra javította a visszakeresési arányt, mivel a fejezetszintű kontextus megőrzése révén a táblázatfejlécek és a kapcsolódó sorok szemantikai összefüggése nem vész el az indexelés során, míg szöveges kérdéseknél mindkét stratégia azonosan teljesített (0,75).

A c) diagram a felhasználói visszajelzéseken alapuló elégedettségi átlagot mutatja 1–5 skálán: a hierarchikus stratégia bevezetésével az átlagos értékelés 3,7-ról 4,6-re nőtt, ami 24%-os relatív javulást jelent.

Az eredmények azt mutatják, hogy a hierarchikus szeletelési stratégia különösen hatékony az intézményi környezetre jellemző, erősen strukturált dokumentumok – tantervek, szabályzatok, mobilitási táblázatok – visszakeresésénél, míg szöveges tartalmak esetén a két módszer teljesítménye egyenértékű.



2. ábra: Szeletelési stratégiák összehasonlítása

4, Retrieval pipeline (visszakeresési folyamat) és válaszgenerálás

A rendszer központi komponense a *Retrieval-Augmented Generation* (Lewis, és mtsai., 2020) pipeline, amely a dokumentum-visszakeresést és a válaszgenerálást egységes folyamatban valósítja meg. Célja, hogy a felhasználói kérdéshez releváns dokumentumszeleteket rendeljen, majd ezek alapján forrásalapú választ generáljon.

A visszakeresési folyamat első lépése a felhasználói kérdés vektoros reprezentációjának előállítása. Ez a vektor az Azure Search hibrid lekérdezésében kerül felhasználásra. A hibrid keresés több módszert kombinál: a BM25-alapú lexikális keresés a kulcsszavas egyezéseket rangsorolja, míg a vektoralapú keresés a kérdés és a dokumentumszeletek szemantikai hasonlóságát vizsgálja (Robertson & Zaragoza, 2009) (OpenAI, 2024). A két megközelítés kombinálása lehetővé teszi a pontos kifejezések és a jelentésalapú egyezések együttes kezelését.

A visszakeresett találatokra szemantikus újrarangsorolás épül. Az Azure Search beépített újrarangsorolója mellett a rendszer saját optimalizációs réteget is alkalmaz, amely figyelembe veszi a relevanciapontszámokat, a dokumentumszeletek metaadatait és a találatok diverzitását. A diverzitási mechanizmus csökkenti annak esélyét, hogy a végső kontextus túl sok, azonos oldalról vagy dokumentumrészből származó szeletet tartalmazzon.

A visszakeresési folyamat dokumentumtípus-specifikus optimalizációkat is támogat. A rendszer képes különbséget tenni szöveges és táblázatos jellegű kérdés-válaszok között, így a strukturált adatokra irányuló lekérdezések más súlyozással kezelhetők. Enterprise környezetben további funkciók is elérhetők, például a lekérdezés-gazdagítás (*query enrichment*), a dinamikus paraméterhangolás és az iteratív visszakeresés. Ezek rövid, többértelmű vagy összetettebb kérdések esetén javíthatják a visszakeresett dokumentumszeletek relevanciáját.

A válaszgenerálás során a kiválasztott dokumentumszeletek kontextusblokként kerülnek a nyelvi modell elé. A rendszer minden szelethez hivatkozási metaadatokat rendel, így a

válaszok nemcsak generált szövegként, hanem konkrét dokumentumforrásokhoz kapcsolva jelennek meg. Ez intézményi környezetben különösen fontos, mert a válaszok pontossága mellett azok visszakövethetősége is alapkövetelmény.

A generált válaszok Azure OpenAI-alapú nyelvi modellekkel készülnek. A rendszer streaming-alapú válaszküldést is támogat, amely a szöveg fokozatos megjelenítésével javítja a felhasználói élményt. A pipeline kialakításának fő célja a hallucinációs kockázat csökkentése: a modell minden esetben explicit dokumentumkontextus alapján állítja elő válaszát.

5. Biztonsági és enterprise mechanizmusok

Intézményi környezetben az AI-asszisztensek esetén alapvető követelmény a dokumentumhozzáférések szabályozása, az erőforrás-felhasználás kontrollja és a felhasználói interakciók monitorozása. A fejlesztett rendszer ezért több enterprise szintű biztonsági mechanizmust tartalmaz.

A hozzáférés-kezelés szerepalapú modellre épül. A felhasználók hallgatói, oktatói, adminisztrátori vagy egyéb jogosultsági szintekhez rendelhetők. A dokumentumhozzáférést hatókör-alapú szűrés egészíti ki, amely biztosítja, hogy a visszakeresési folyamat kizárólag a felhasználó számára elérhető dokumentumállományban keressen. A keresési szűrők whitelist (engedélyezési lista)-alapú validációt használnak, így csak előre engedélyezett mezőkön keresztül hajthatók végre.

A hitelesítés elsődlegesen Microsoft Entra ID-alapú egyszeri bejelentkezéssel történik. A rendszer emellett API-alapú hozzáférést is támogat, amely JWT tokenekkel vagy partnerintegráció esetén API-kulccsal működhet. A prompt injection jellegű mintázatok felismerésére bemeneti szűrés szolgál, amely naplózza a gyanús lekérdezéseket.

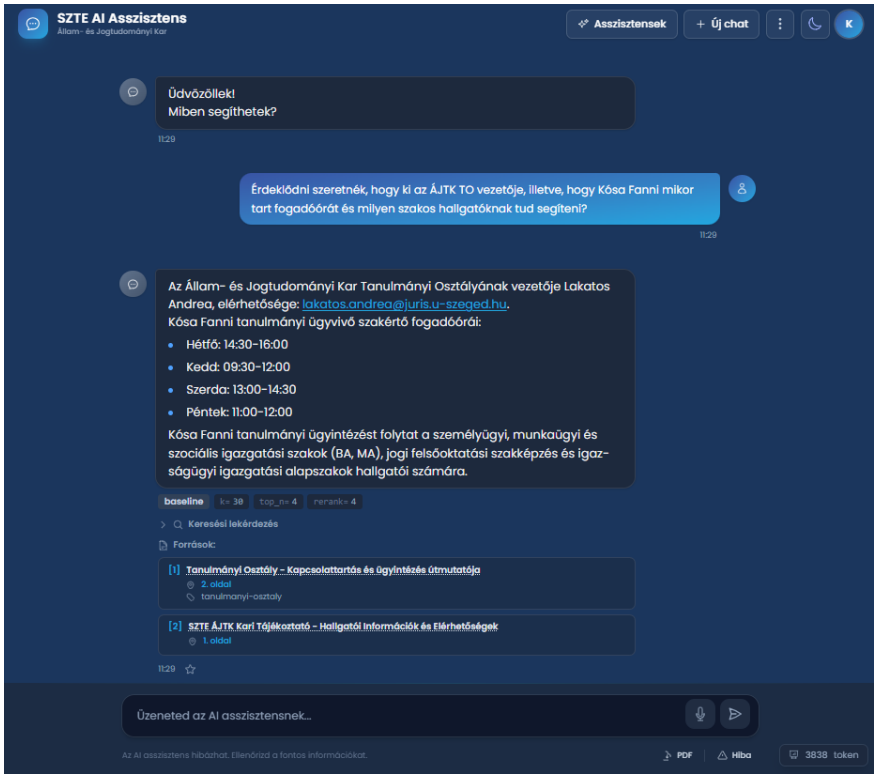
A működési stabilitást tokenkvóta-alapú erőforrás-szabályozás támogatja. A kvóták globális, szerepköri vagy felhasználói szinten határozhatók meg, így megelőzhető a túlzott modellhasználat. Az enterprise pipeline része a visszakeresési nyomkövetés is, amely a keresési paramétereket, iterációkat, időigényt és tokenfelhasználást rögzíti. Ez támogatja a teljesítményelemzést és a későbbi optimalizációt.

A rendszer naplózás-szanitizálási megoldásokat is alkalmaz, amelyek eltávolítják a naplókból az érzékeny adatok egy részét, például e-mail-címeket vagy telefonszámokat. Az adminisztratív műveletek auditnaplózása biztosítja a dokumentumkezelési, jogosultságkezelési és konfigurációs változások visszakövethetőségét. Ezek a mechanizmusok együttesen támogatják a biztonságos és kontrollált intézményi működést.

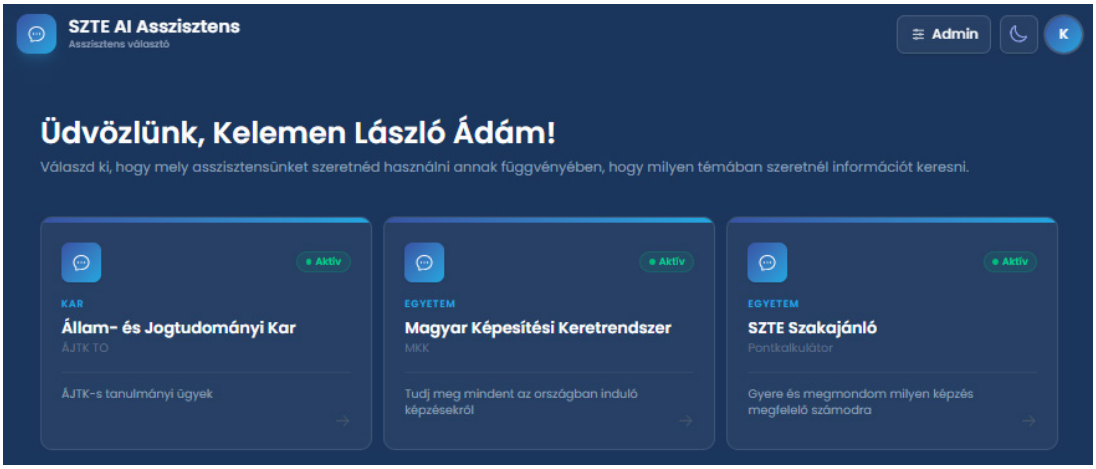
6. A rendszer felhasználói felülete

Az alábbi ábrák a fejlesztett mesterséges intelligencia asszisztens keretrendszer egyes felhasználói felületi komponenseit és működési példáit mutatják be. A képernyőképek szemléltetik a természetes nyelvű lekérdezések kezelését, a dokumentumalapú válaszgenerálást és az intézményi környezetben alkalmazott funkcionalitásokat.

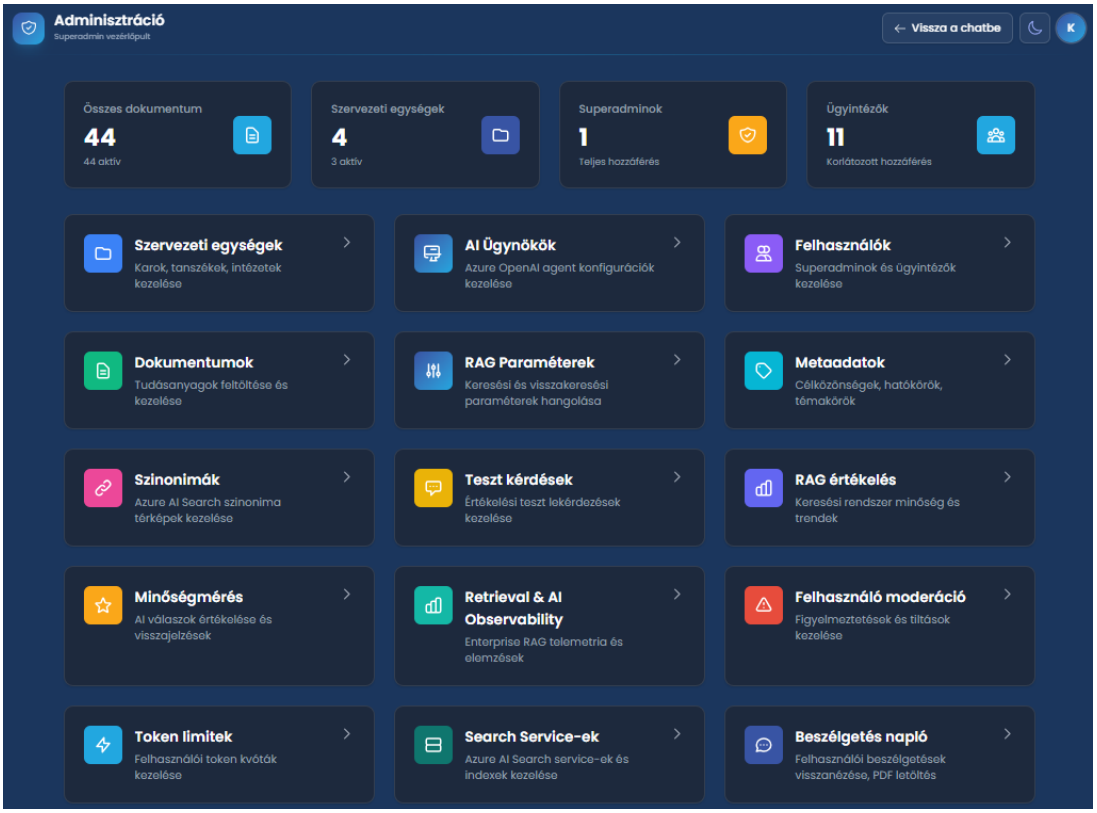
3. ábra: Belépési felület



4. ábra: Dokumentumalapú válaszgenerálás és forráshivatkozások a felhasználói felületen



5. ábra: Asszisztensválasztó felület



6. ábra: Az AI-asszisztens adminisztrációs és visszakeresési menedzsmentfelülete

7. Következtetések

A tanulmány egy felhőalapú *Retrieval-Augmented Generation* architektúrára épülő mesterséges intelligencia asszisztens keretrendszert mutatott be, amely intézményi dokumentumkezelési és információ-visszakeresési feladatokat támogat. A rendszer hibrid keresést, vektoralapú szemantikus visszakeresést, szemantikus újrarangsorolást és többféle szeletelési stratégiát alkalmaz.

A fejlesztés egyik fontos eredménye, hogy a rendszer dokumentumfeldolgozási lánca képes eltérő szerkezetű intézményi dokumentumok kezelésére. A *paragraph-first chunking* a paragrafusszintű szemantikai koherenciát, míg a *hierarchical chunking* a fejezetszintű kontextus megőrzését támogatja. A visszakeresési folyamat enterprise optimalizációi – például a dinamikus paraméterhangolás és az iteratív visszakeresés – tovább javíthatják a releváns dokumentumszeletek kiválasztását.

A biztonsági komponensek, köztük a szerepalapú hozzáférés-kezelés (RBAC), a hatókör-alapú dokumentumszűrés, a kvótakezelés és a visszakeresési nyomkövetés alkalmassá teszik az architektúrát intézményi környezetben történő használatra. Az eredmények alapján a RAG megközelítés hatékonyan alkalmazható egyetemi dokumentumállományok

természetes nyelvű lekérdezésére, miközben csökkenti a generatív modellek hallucinációs kockázatát és javítja a válaszok visszakövethetőségét.

A jövőbeni fejlesztési irányok közé tartozik a visszakeresési folyamatok további optimalizálása és a találatok relevanciájának növelése, különös tekintettel a különböző szeletelési stratégiák kvantitatív összehasonlítására és a visszakeresési paraméterek automatikus adaptációjára. A fejlesztés további célja az egyetemi tanulmányi rendszerekkel történő mélyebb integráció kialakítása, amely lehetővé teheti adminisztratív és tanulmányi folyamatok intelligens támogatását. Emellett a keretrendszer új funkcionális iránya a hagyományos kérdés-válasz alapú működés kiterjesztése párbeszédorientált, ajánlórendszer jellegű interakciókra. Ennek egyik példája a fejlesztés alatt álló szakajánló asszisztens, amely természetes nyelvű párbeszéd segítségével képes a felhasználói preferenciák és érdeklődési körök alapján releváns képzési ajánlásokat megfogalmazni.

Irodalomjegyzék

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., . . . Kiela, D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Advances in Neural Information Processing Systems 33. 33.* NeurIPS. Forrás: https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html?utm_source=chatgpt.com

OpenAI. (2024). *Vector embeddings*. Letöltés dátuma: 2026. 05 07, forrás: OpenAI API Documentation: https://platform.openai.com/docs/guides/embeddings

Robertson, S., & Zaragoza, H. (2009). The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends in Information Retrieval, 3*. Forrás: https://dl.acm.org/doi/10.1561/1500000019

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., . . . Polosukhin, I. (2017). Attention Is All You Need. *Advances in Neural Information Processing Systems 30. 30.* Long Beach, California: NeurIPS. Forrás: https://papers.nips.cc/paper/7181-attention-is-all-you-need