https://doi.org/10.60030/ALMAMS.c202301lqckfp

Tudásmenedzsment a Nagy Nyelvmodellek korában

Szabados Levente

Frankfurt School of Finance and Management
Soproni Egyetem
szabados@int.fs.de
ORCID: 0009-0003-6046-9361

Absztrakt

A mesterséges intelligencia (AI) fejlődése és különösen a nagy nyelvmodellek (LLM-ek) elterjedése alapvetően változtatja meg a vállalati tudás-menedzsment (KM) lehetőségeit. A tanulmány célja, hogy megvizsgálja, hogyan használhatók az LLM-ek a vállalati tudás hatékony kezelésében és felhasználásában. Az LLM-ek által kínált előnyök, mint a nagymértékű szöveges adatbázisokból való tanulás és az új információk generálása, jelentős potenciált hordoznak, ugyanakkor a hallucinációk és a pontatlan következtetések jelentős akadályokat jelentenek alkalmazásukban. A tanulmány különböző technikákat tárgyal, amelyekkel ezen kihívások mérsékelhetők, mint például a „Chain-of-Thought" eljárás, a finomhangolás, valamint a „Retrieval Augmented Generation" (RAG) módszerek. Végezetül a tanulmány empirikus adatokat elemez a szoftverfejlesztők és vállalati döntéshozók körében végzett felmérések alapján, bemutatva az LLM-ek tudásmenedzsmentben történő alkalmazásának lehetőségeit, kihívásait és várható hatásait.

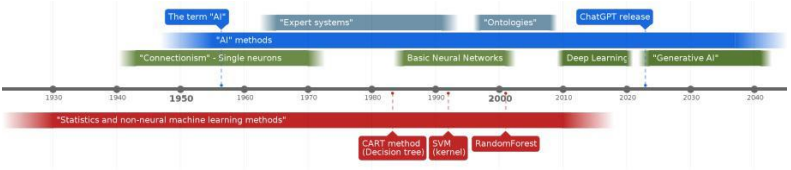
Kulcsszavak: *mesterséges intelligencia (AI), nagy nyelvmodellek (LLM-ek), vállalati tudásmenedzsment (KM)*

Háttér és motiváció

A közgazdasági kutatás utóbbi néhány évében különösen sok eredmény (például: [29], [25], [10]) látszik alátámasztani a tényt, miszerint a vállalati tudás és különösen annak hatékony felhasználása kulcs fontosságú versenyképességi tényezőt jelent a szervezetek számára (a tudástőke mérése és annak versenyképességre gyakorolt hatása kapcsán lásd még [1]), így kiemelten fontos lehet vizsgálni, hogy a legújabb informatikai megoldások, különösen az utóbbi években látványos fejlődést és elterjedést mutató mesterséges intelligencia (AI) megoldások milyen módon képesek elősegíteni a vállalati tudásmenedzsmentet, s így miként képesek (a közvetlen feladat automatizáláson túl) hozzájárulni a szervezeti versenyképességhez.

A mesterséges intelligencia új korszakának háttere és technológiai fejlődése

A mesterséges intelligencia (AI) technológia fejlődése jelentős utat járt be azóta, hogy 1955-ben bevezették a fogalmat [28]. Az AI kutatása különböző fázisokon ment keresztül, amelyek mindegyike egy adott technológia vagy kutatási irány dominanciáját jelölte.



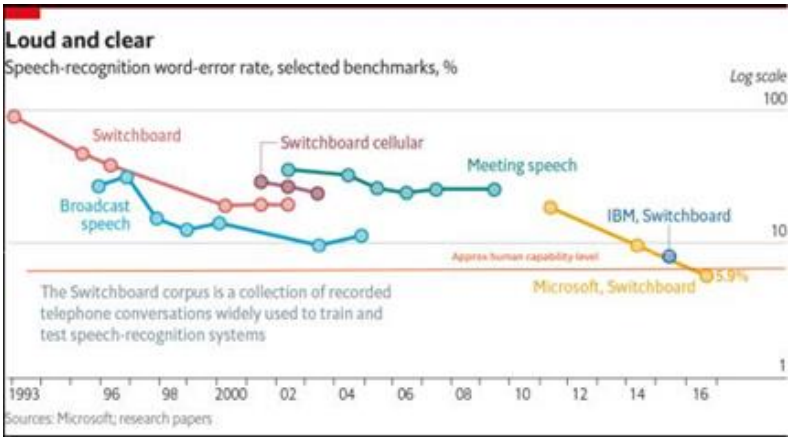
1. ábra. Az AI fejlődésének idővonala

A klasszikus AI kutatási területek, mint az automatizált beszédfelismerés vagy a képosztályozás, évtizedekig folyamatos előrehaladást mutattak, azonban ezen fejlődés jelentős mérnöki erőfeszítést igényelt. Az "AlexNet" példája [24], amely az első sikeres mélytanuló modell ("mély neurális hálós tanulás" avagy "Deep Learning" - fogalom eredetéhez lásd még: [6], [12]) volt, amely felülmúlta a hagyományos (nem neurális háló alapú) statisztikai modelleket, bemutatja az "end-to-end learning" paradigmaváltást. Ez a megközelítés lehetővé tette, hogy nagyméretű adatbázisokon (például az ImageNet-en [8]) alkalmazzunk modelleket anélkül, hogy szükség lenne az emberi szakértők által manuálisan tervezett jellemzőkre, így gyakorlatilag a mérnöki órák számát cseréltük le a számítástechnikai erőforrásokra.

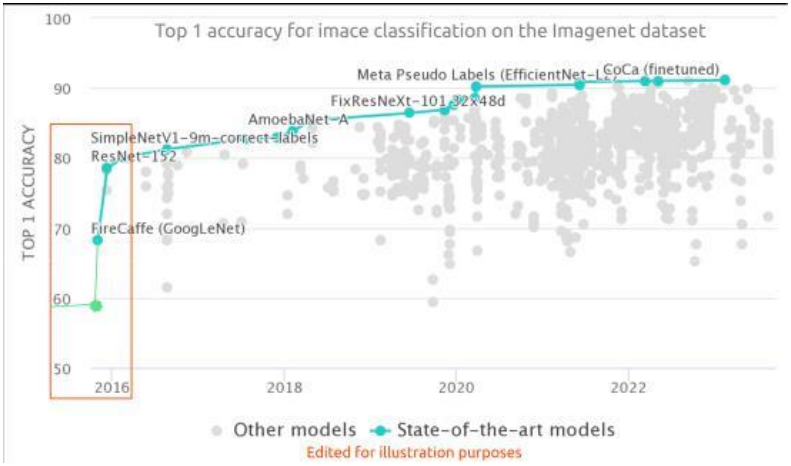
Figure 1: Trends in $n = 121$ milestone ML models between 1952 and 2022. We distinguish three eras. Notice the change of slope circa 2010, matching the advent of Deep Learning; and the emergence of a new large-scale trend in late 2015.

2. ábra. A legkorszerűbb AI modellek számítási igénye idővel [37]

Ez a változás - a számítástechnikai erőforrások széles körű elérhetőségével és a nagyobb adatbázisok megjelenésével párosulva - áttörést eredményezett az elért teljesítményben több lényeges területen, mint a hangátírás és a képfelismerés vagy szövegfeldolgozás.



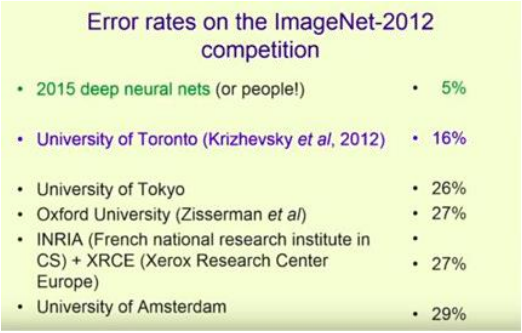
3.ábra. Az automatizált beszédfelismerés története [15] (módosítva az emberi szintű teljesítmény hozzávetőleges megjelenítésére)



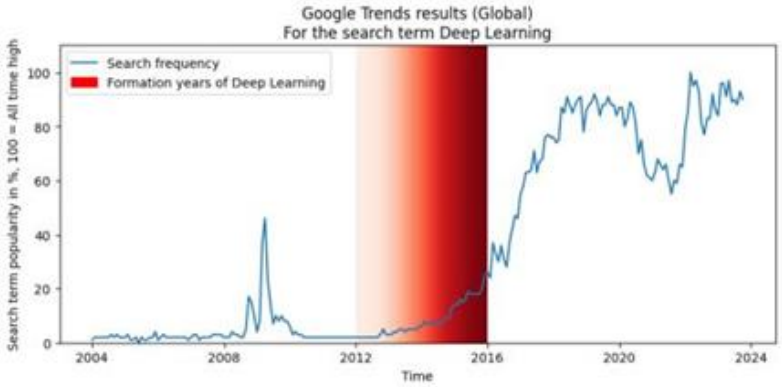
4. ábra. Képosztályozási teljesítmény ImageNet adatszetten

Amint az AI egyik "alapító atyja", a Turing-díjas Geoffrey Hinton nyilvános előadásában kiemelte [17], a 90-es évek végének elméleti előrelépései csak akkor hoztak gyümölcsöt, amikor elegendő adat és számítási kapacitás vált elérhetővé egy megfelelő modellstruktúra (azaz a mélytanulás) számára. A korábbi modellezési architektúrák nem élveztek közvetlen

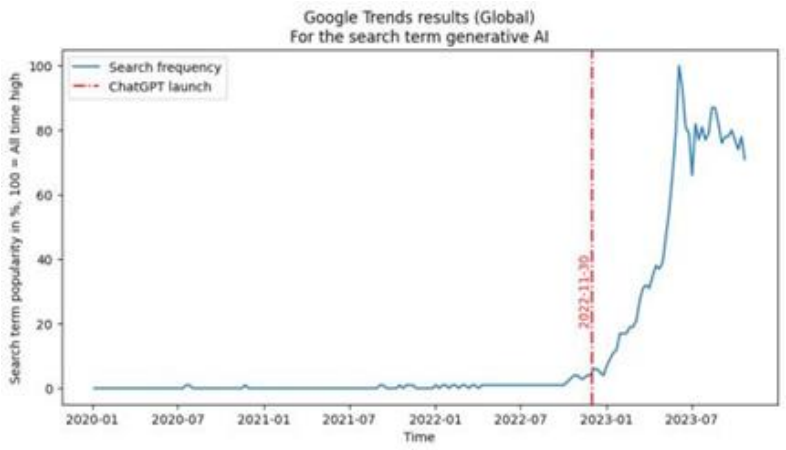
előnyt az adatok és a számítási kapacitás növekedéséből. Ahogyan Hinton megjegyzései illusztrálják, a képosztályozás területén évtizedekig tapasztalt alacsonyabb teljesítmény (26%-os hibaarány) 2012-ben az első jól képzett mélytanuló modellel majdnem felére csökkent, majd körülbelül 3 év alatt elérte az emberi szintű teljesítményt. Amint ez a jelentős előrelépés és a teljesítmény gyors növekedése széles körben ismertté vált (ahogy ezt például a "Deep Learning" kifejezésre vonatkozó Google keresési trendek is mutatják), az AI technológia egy újabb szakaszba lépett. A technológia fejlődése a 2020-as évek elején új magaslatokat ért el a generatív AI rendszerek megjelenésével, amelyek képesek szöveges vagy képi válaszokat generálni, és amelyekre a hétköznapi nyelv a „generatív AI' kifejezést használja.



5. ábra. Geoffrey Hinton előadása a mélytanulás történetéről [17]



6. ábra. Google Trends keresés: Deep Learning [13]



7. ábra. Google Trends keresés: generative AI [14]

Ezeket a rendszereket a tudományos közösségben gyakran "megalapozó modellek" (Foundational Models [2]) névvel illetik, mivel ezek a modellek nemcsak hogy pontosabbá váltak, hanem alkalmazhatósági körük is robbanásszerűen szélesedett, azaz bizonyos értelemben általános problémamegoldó modellekként funkcionálnak, nem szűkülve egyetlen specifikus alkalmazási területre sem, valamint szabad szöveges formában lehet velük interakcióba lépni, mintegy "instrukciókat" adva, így tehát a használatukhoz szükséges technológiai küszöb is jelentősen csökkent, a felhasználók széles köre hozzáfér teljesítményükhöz specialista tudás nélkül is. A generatív AI rendszerek fejlődése, többek közt a diffúzió alapú képgeneráló modellek [18] és a nagy nyelvi modellek (Large Language Model, LLM) kombinációja, a multimodalitás koncepciója alatt új lehetőségeket és kihívásokat hozott az AI területére. Ez a technológiai ugrás jelentős figyelmet keltett mind a tudományos, mind a gyakorlati alkalmazások terén.

Nagy nyelvmodellek és tudás

Témánk szempontjából különösen kiemelendő az LLM-ek szerepe, mivel roppant méretű szöveges korpuszokon tanítva tulajdonképpen tekinthetők úgy, mint az emberi tudás széles körének "tömörített" tudásbázisai (lásd még: [7]), melyek szabad szöveges formában "lekérdezhetők", a "transzfer tanulás" ([32]) és valamint később "promptolás" ([3]) segítségével alkalmasak új, korábban nem látott kérdések megválaszolására. Bár a fenti érvelés egykönnyen arra engedne következtetni, hogy az LLM-ek már önmagukban, extenzív tréningjük, és a felhasznált roppant méretű tanító adatbázisok lefedése miatt értékes tudásforrásnak minősülhetnek, így jelentősen támogathatják a vállalati tudásmenedzsmentet, sajnálatos módon a gyakorlat evvel homlokegyenest ellentétes tapasztalatokat mutat. A nagy

nyelvi modellek esetében, főként az "instrukciós finomhangolás" (lásd.: [30]) miatt, mely a puszta nyelvi képességeken túl egyfajta viselkedéses elvárást teremt a modellek felé, fokozottan fellép a "hallucináció" jelensége, mely során a modell mindenáron törekszik az instrukciók teljesítésére, az általa generált szövegek tényszerűségének teljes figyelmen kívül hagyásával. E jelenség, melyet [21] részletesen tanulmányozott, kifejezett akadálya az LLM-ek tudásmenedzsment területén való felhasználásának oly mértékig, hogy önmagukban nézve (tehát egyéb kiegészítő technikákat nem felhasználva) kijelenthető, hogy az LLM-ek tudásforrásnak alkalmatlanok (lásd még: [42]). Kérdés tehát, hogy milyen eljárások alkalmazhatóak az LLM-ek tudásmenedzsment területen történő hasznosításában?

Lehetséges technikák LLM-ek hasznosítására a tudásmenedzsmentben

Általános keretek

A hipotetikus kontextus, melyben az LLM-ek vállalati tudásmenedzsment kontextusban történő alkalmazását vizsgálni kívánjuk, a következő:

A vállalat valamely üzleti folyamata során információ igény keletkezik vagy egy külső szereplő (például ügyfél), vagy egy belső szereplő (munkatárs) irányából, melyet automatizált formában, például egy kérdés-válasz keretében szeretnénk kielégíteni. Feltesszük továbbá, hogy az információ igény a szervezeten belül expliciten, például írásos dokumentumok formájában rendelkezésre álló tudásra, és a belőle levonható közvetlen következtetésekre korlátozódik. Kérdés, hogy milyen formában tudunk LLM megoldásokat használni ezen információ igények kielégítésére az esetben, ha a hallucináció problémáját figyelembe vesszük?

Problémák alábontása

Ahhoz, hogy a fent leírt praktikus alkalmazási mód lehetővé váljon, meg kell különböztetnünk az LLM hallucinációk két alapvető formáját, a következtetési tévedéseket és a tudásdeficitból adódó fikciókat.

Míg az első esetben arról van szó, hogy az adott nyelvi modell számára - a tréning adatból következően, vagy a kérdés bemeneti kontextusában ("prompt"-jában) - rendelkezésre áll minden információ a helyes következtetés levonására, mégis téves választ ad, míg a második esetben a modell tréning során vagy a kérdésben nem kapott elegendő tényszerű információt a válaszhoz, így kénytelen a legvalószínűbb szöveget tulajdonképp véletlenszerűen "megtippelni", mely során szinte bizonyosan tárgyi tévedésekbe bocsátkozik. E két problémát, és a rá adott válaszokat külön kell vizsgálnunk.

Következtetési hibák korrekciója

A következtetési hibák korrekciója terén szerencsére idejekorán történtek előrelépések, mivel az úgynevezett "Chain-of-Thought" technika [41] , és a hozzá tartozó, fundamentálisnak tűnő jelenség igen ígéretes irányt jelent.

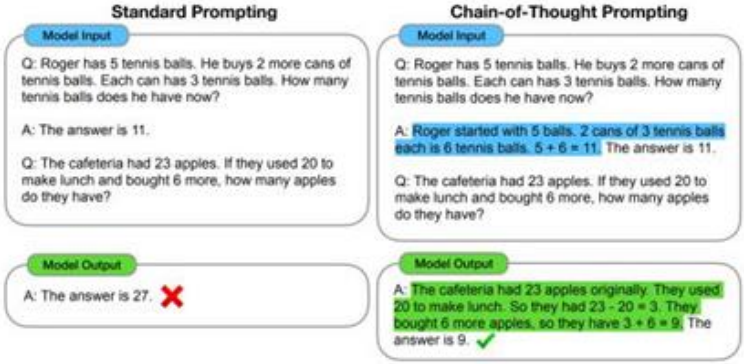


Figure 1: Chain-of-thought prompting enables large language models to tackle complex arithmetic, commonsense, and symbolic reasoning tasks. Chain-of-thought reasoning processes are highlighted.

8. ábra. "Chain-of-Thought" technika demonstrációja az eredeti cikkben [41]

A technika lényege, hogy az LLM-et megfelelő utasításokkal arra késztetjük, hogy lépésről lépésre bontsa fel válaszait, mintegy megtervezve majd végrehajtva a következtetési láncot, mely egy adott probléma megoldásához vezet, tulajdonképp saját kimenetét (és az ehhez kapcsolható komputációs "büdzsét") használva "jegyzettömbként" a következtetés jó minőségű végrehajtására.

A fenti technika egyben egy komoly kutatási terület kezdeti pontját is jelenti, mely mára számos hatékony eljárást foglal magába az LLM-ek következtetési képességeinek javítása érdekében. A terület legújabb eredményeiről lásd még: [36] és [4]

Ezek alapján kijelenthetjük, hogy amennyiben a megfelelő információ rendelkezésre áll, több módszert is bevethetünk arra, hogy az LLM válaszaiban megjelenő következtetési hibákat csökkentsük.

Mi azonban a helyzet akkor, ha tényszerű "ismerethiányról" van szó, azaz a tulajdonképpeni kérdésről: A saját, különleges vállalati tudásom, mely nem volt része az eredeti LLM tréning anyagnak, hogyan kerül a modellbe?

Finomhangolás

Bár a technika gyökerei jóval korábbra nyúlnak vissza, talán elsőként az ULMFiT modell [19] irányította rá a fókuszt a "finomhangolás" kérdésére. A megközelítés lényege, hogy a nagy méretű szövegkorpuszon korábban "előtanított" mély neurális hálós nyelvmodelleket az adott kontextus, jelen esetben szervezet számára elérhető szöveganyag felhasználásával "tovább tanítjuk", így a modellnek lehetősége van mind a vállalati tudást, mind a kommunikáció stílusát elsajátítani.

Az első és legkomolyabb kihívás ebben a megközelítésben - az elérhető vállalati adatbázis méretén túl - a jelentős számítási igény, mellyel ténylegesen nagy nyelv-modelleket finomhangolni lehet. Ez gyakran meghaladja a szervezetek által akár felhős környezetben bérelhető kapacitásokat is. A szükséges számítási kapacitás mérséklésére létrehozott megoldások, mint például a LoRA [20] és QLoRA [9] módszerek ugyan jelentős mértékben mérséklik a szükséges befektetést, ám az – különösen a módszerekhez szükséges speciális szakértelem figyelmbevételével együtt – erőteljes befektetést igényelnek.

Sajnálatos módon számos megfigyelés támasztja alá a tényt, miszerint a "finomhangolás" egyben az LLM korábbi, előtanulási szakaszában elsajátított tudásának sérüléséhez vezet, sőt, egyes eredmények (például: [23]) szerint erős lineáris kapcsolat áll fenn a finomhangolás mértéke és a "felejtés" között. A tartalmi "tanulás" szempontjából pedig komoly érvek (lásd [31]) szólnak amellett, hogy a "finomhangolás" speciális tudás elsajátítására nem feltétlenül alkalmas.

A finomhangolás és transzfer tanulás kapcsán érdemes megemlítenünk a kérdést, miszerint nem lehet-e egy adott szervezeti tudást finomhangolás útján "tárolni" egy LLM-ben, így egyfajta tacit tudásként átvinni más szervezeti kontextusokba. Bár ez elméletben lehetséges, fontos belátnunk, hogy mivel a tudás kiindulópontja explicit tudást tároló dokumentum halmaz volt (a tréning korpusz formájában), ezért ez a megközelítés előbb explicit tudást tacittá tesz, majd azt viszi át szervezeti határokon. Mint ilyen, egyfajta veszteséges tömörítésnek minősül, így nem tekinthető preferáltnak az explicit tudást átvivő dokumentum transzferhez képest.

Mindezek alapján a finomhangolás a vállalati tudás, az adott szervezeti információk LLM-ek felé történő tanításának messzemenőkig nem optimális formája, és mint ilyen, erre a célra kerülendő, sokkal inkább alkalmas valamiféle írásos formában megnyilvánuló "skill", mint kommunikációs stílus, hangvétel, viselkedés LLM-ek felé való átadására, feltéve hogy ez a tréning korpuszban megfelelően tükröződik. Az explicit tudás LLM-ek felé történő átadására tehát egyéb megoldásokat kell keresnünk.

Retrieval Augmented Generation (RAG)

A speciális, adott esetben csak szervezeten belül elérhető tudás nyelvi modellekkel való "összekötése" már igen korán felmerült, és az azóta megalapozónak tekinthető cikk [27] fogalmait követve a keresőrendszerekkel kiterjesztett nyelvi generálás (Retrieval Augmented Generation - RAG) a kutatás és gyakorlati alkalmazás fókuszpontjába került. Az eljárás lényege, hogy egy testre szabott, gyakran szemantikus feldolgozást is végző keresőrendszert integrálunk az LLM megoldásokkal oly módon, hogy a beérkező kérdésekhez előbb keresést futtatunk a vállalati tudásbázisban, majd annak releváns találatait mintegy "horgony-ként" ("anchor") vagy "kontextusként" ("context") adjuk át a kérdéssel együtt az LLM-nek, így az tényszerű válaszait az eredeti információforrásokhoz tudja kötni, minimalizálva a hallucináció esélyét.

A módszer sikerét mutatja, hogy már önmagában is jelentős kutatási területté vált (lásd még pl: [22]), és az LLM-ek vállalati munkafolyamatokba való beillesztésének gyakorlatilag de facto megközelítésének tekinthető.

Várható hatások

Miután fentebb körvonalazott megoldások, különösen a RAG megközelítés jelentős sikereket ért el az LLM-ek vállalati tudásmenedzsmentbe való beilleszthetősége terén, térjünk át vizsgálatunk másik tárgyára: Mekkora potenciált jelent ez a technológia a gyakorlatban, milyen mértékű hatások várhatóak, amennyiben jelentősen elterjed, illetve mik az alkalmazásának főbb kihívásai? Mindezen kérdések tisztázásához empirikus adatok elemzéséhez nyúlunk, még-hozzá két forrásból:

Első forrásunk, a szoftver iparágban domináns információ forrásnak tekinthető "Stack Overflow" portál által évenként elvégzett közvélemény kutatás ("Stack Overflow Developer Survey 2023" [38]), mely bár iparági értelemben korlátozott (hisz szoftverfejlesztőket kérdez meg), viszont a válaszadók impresszív száma (közel 90e fő világszerte) miatt mérvadó adatforrásnak tekinthető.

Második forrásunk a Trust My AI és a Neuron Solutions Kft. által végzett saját vezetői közvélemény kutatás (továbbiakban "TrustMyAI kutatás"), mely célzottan a régiónkban tevékenykedő vállalati döntéshozókat (n=150) kérdezett meg az AI megoldások, különösként az LLM-ek alkalmazása és a vele kapcsolatos megbízhatósági kérdések tekintetében.

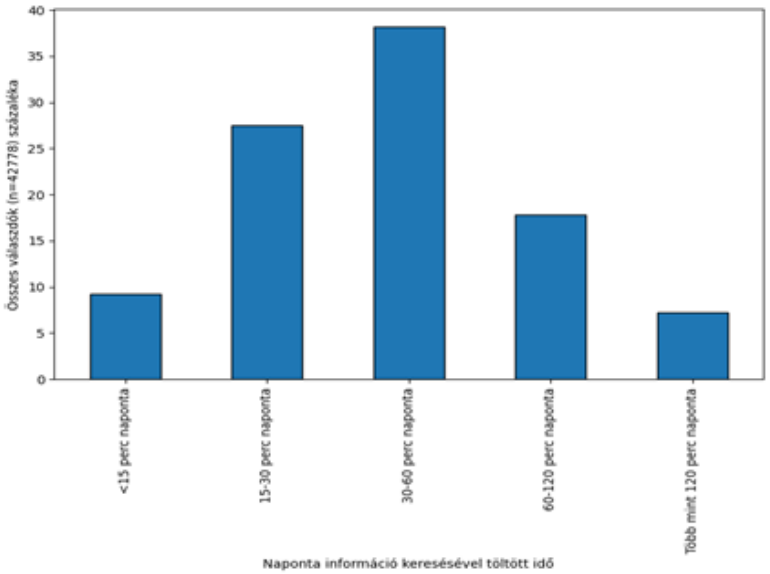
Mindkét adatforrás a kérdések tágabb körét taglalja, jelen elemzésünkben csupán azok egyes releváns kérdéseit dolgozzuk fel.

Igény

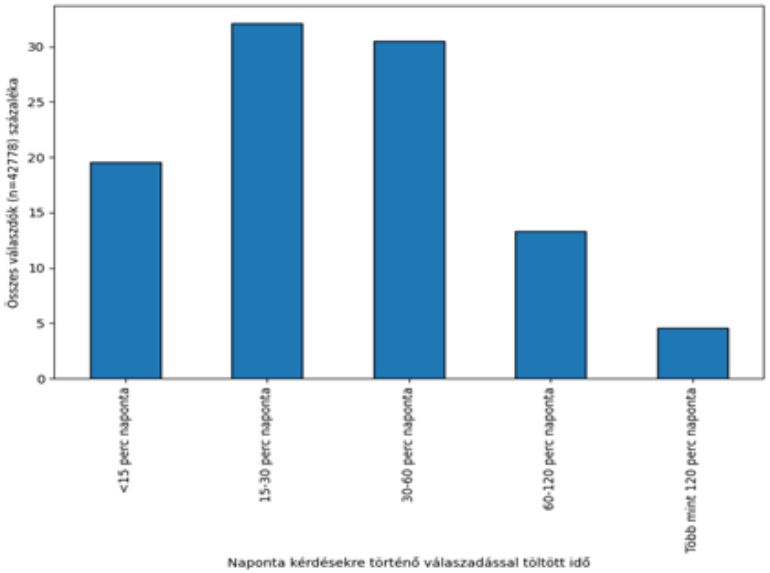
Elsőként vizsgáljuk meg, hogy egy tudás intenzív munkakörben - mint például a programozás - milyen potenciál rejlik a tudásmenedzsment eszközök hatékonyabbá tételében?

Annak ellenére, hogy fő tevékenységükként a programozók új forráskód létre-hozását jelölik meg, igen instruktív megfigyelni, hogy saját bevallásuk (Stack Overflow survey) szerint naponta mennyi időt töltenek információk keresésével és kérdések megválaszolásával.

Még ha csupán a megjelölt idősávok átlagértékeivel számolunk is, elmondható, hogy egy átlagos programozó a napjából közel 50 percet információ kereséssel, 40 percet pedig kérdések megválaszolásával tölt, ami a munkaidő több mint 18%-a! Ez az adat élesen kiemeli, hogy ha LLM megoldásokkal képesek vagyunk csökkenteni az információhoz való hozzájutás idejét, igen jelentős termelékenység növekedés érhető el. Megjegyzendő, hogy a legfrissebb, 2024-es Stack Overflow felmérés[39] adatai szinte tökéletesen megegyeznek a 2023-as mérésekkel ebben a tekintetben, így a jelentős potenciálon túl még nem árulkodik az adat annak komoly kiaknázásáról. Felmerülhet tehát a kérdés, hogy pontosan mennyire élnek már evvel a lehetőséggel a szervezetek a gyakorlatban?



9. ábra. A Stack Overflow felmérés [38] alapján naponta mennyi időt töltenek programozók válaszok keresésével



10.ábra. A Stack Overflow felmérés [38] alapján naponta mennyi időt töltenek programozók kérdések megválaszolásával

Alkalmazás

A fenti becslések csupán a lehetőségekről szólnak. Kérdés, hogy van-e bármiféle evidenciánk arról, hogy evvel a lehetőséggel a gyakorlatban élnek is? Ehhez ismét segítségül hívhatjuk a fenti adatbázist.

A megkérdezett szenior (legalább 3 éve professzionálisan a szakmában dolgozó) programozók (n=57240) közül 10.61% említette, hogy egy kódbázissal való ismerkedéshez, 17.01% pedig a hibajavításhoz és segítség kéréshez használ jelenleg is már AI (LLM) alapú technológiát, az ezt csupán tervezők esetében a két szám 19.74% és 17.10%. (Forrás: Stack Overflow survey [38], saját elemzés)

Emellett a területen végzett kvalitatív kutatás [26] során is fény derült rá, hogy a megkérdezett szenior programozók igen gyakran említették: már most is jelentős mértékben támaszkodnak LLM alapú rendszerekre mintegy "orákulumként", azaz a nagy mennyiségű dokumentációval vagy ismeretlen kódbázissal való ismerkedés terén mintegy kérdés megválaszoló és tanulást segítő eszközként. Ez jól korroborálja a feltevést, hogy a fent említett időmegtakarítási potenciál a gyakorlatban is megvalósulhat.

Akadályok

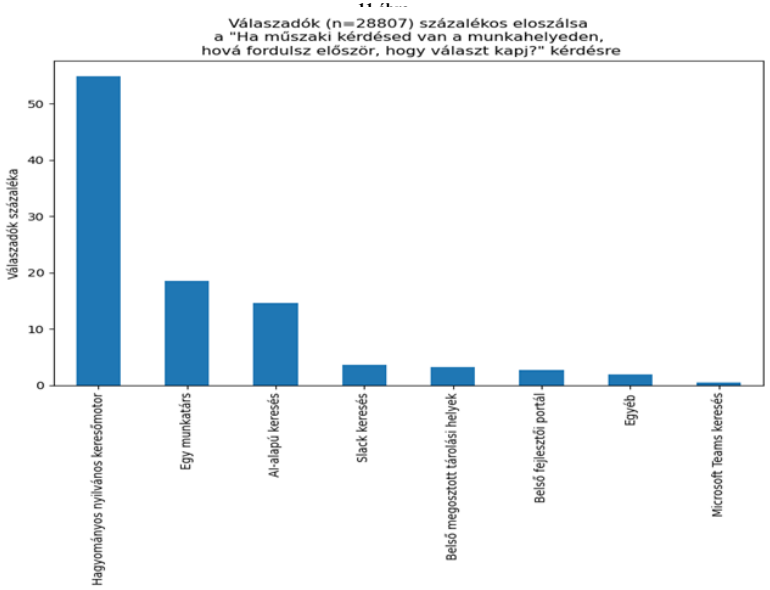
Mindezek mellett erős akadályozó tényezők is fellépnek.

Az LLM megoldások biztonságos és felelős használata még továbbra sem megoldott. Ahogy azt a Conversica által végzett felmérés [5] is mutatja, a mindössze a válaszadók 6%-ának van már bevezetett irányelve, ráadásul a következő 12 hónapban mesterséges intelligenciát bevezetni tervező vállalatok közül mindössze minden huszadiknak van irányelve erre vonatkozóan. A dolgozók tehát nem tudják, mit szabad és hogyan biztonságos LLM megoldásokat használni. (Ez szintén tükröződött a szenior programozókkal végzett mélyinterjúimban [26].)

Amikor a Stack Overflow 2024-es felmérésében[39] azt vizsgálták, hogy milyen forrást is kérdezne meg leginkább egy programozó, ha problémája van, azt találták, hogy bár az AI alapú megoldások (fizetős és nem fizetős formában) összességében már a a szakemberek 14.57%-ának szemében tekinthetők elsődleges forrásnak (evvel igen komolyan megközelítve a második helyen álló "munkatárs" kategóriát, a maga 18.48%-ával), mégis egyelőre, valószínűleg az elterjedtség / ismertség, vállalati irányelv szerinti elérhetőség, illetve megszokás miatt még továbbra is a nyílt Internetes keresők dominálnak. Ez véleményem szerint erőteljesen változni fog, ha az AI megoldásokba vetett bizalom megnövekszik.

Mitől lesz használható

Fontos kiemelni a bizalom / megbízhatóság kérdésének elsődlegességét. Mivel a hallucináció jelensége erőteljesen befolyásolja az LLM alapú rendszerek megbízhatóságát, az egyik legfontosabb feladat - amint azt már fentebb a RAG



11. ábra A Stack Overflow 2024 felmérés [39] alapján mit tekintenek a programozó szakemberek elsődleges forrásnak

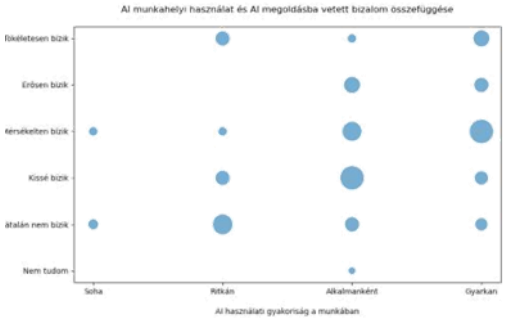
megoldások tárgyalásakor is kiemeltük - az LLM alapú megoldások tényszerű lehorgonyzásában, és annak minőségében rejlik.

Ezt a felismerést támasztja alá szűkebb területen az általam korábban végzett[26] kvalitatív vizsgálat, mely során a bizalom megteremtésének feladata volt az egyik elsődleges kihívás a szenior szoftverfejlesztők számára az LLM alapú megoldások munkafolyamataikba való beillesztése esetében.

A szűkebb körű és kvalitatív evidenciák mellett azonban szélesebb lefedésű kvantitatív adataink is vannak (köszönhetően a Trust My AI és a Neuron Solutions által végzett felmérésnek) arról, hogy az üzleti döntéshozók (n=150) döntő többsége a bizalom kérdését, és az annak megteremtésére tett technológiai erőfeszítéseket (például RAG rendszerek metodikus mérését és benchmarking utáni továbbfejlesztését) mennyire tekintik kiemeltnek a sikeres vállalati bevezetés esetében.

Az adatok numerikus elemzése alapján szignifikáns korreláció (Spearman R: 0.371, P: 0.00001) figyelhető meg a döntéshozók AI megoldásokba vetett bizalma és az alkalmazás gyakorisága között. (Ez egybecseng [40] eredményeivel.)

Megjegyzendő, hogy itt nem csupán egyfajta passzív bizalomról van szó, hanem kifejezetten arra irányult a kérdés, hogy az adott szervezet mennyit tesz technikai és folyamati eszközökkel a megbízhatóság érdekében. Ez jól egybecseng a kutatásban fontos témaként megjelenő "RAG audit" módszertanok (például: [11], [35], [33], [16], [43], [34]) térnyerésével, ezek alkalmazása esetén pozitív hatással kecsegtet.



12.ábra. A TrustMyAI és Neuros Solutions felmérés alapján milyen összefüggés van az üzleti döntéshozók szerint az AI megoldásokba vetett bizalom és azok alkalmazási gyakorisága közt

Következtetések

A fentebb kifejtettek fényében összegezve kijelenthetjük, hogy az LLM modellek alkalmazása jelentős mértékben segítheti a vállalati tudás hatékony alkalmazását, s mint ilyen, közvetlenül befolyásolhatja a versenyképességet, ám csupán néhány alapvetően fontos feltétel teljesülése esetén:

Egyrészről a nagy nyelvmodellek puszta alkalmazásán túl szükség van jó minőségű RAG keretrendszer létrehozására, ezen túl pedig olyan módszerek és eszközök alkalmazására, melyek a rendszer válaszainak minőségbiztosításán keresztül hozzájárulnak a megoldásba vetett bizalom kialakulásához.

Ezek nélkül a feltételek nélkül az LLM-ek jelentette ígéret csupán ígéret marad.

Hivatkozások

[1] György Boda és tsai. Sodródásból feljebb lépés. Tudásalapú versenyképesség. Akadémiai Kiadó Zrt., 2024. isbn: 9789636640231. url: https://m2.mtmt.hu/api/publication/35090812. https://doi.org/10.1556/9789636640231

[2] Rishi Bommasani és tsai. On the Opportunities and Risks of Foundation Models. 2022. arXiv: 2108.07258 [cs.LG].

[3] Tom B. Brown és tsai. Language Models are Few-Shot Learners. 2020. arXiv: 2005.14165 [cs.CL]. url: https://arxiv.org/abs/2005.14165.

[4] Zheng Chu és tsai. Navigate through Enigmatic Labyrinth A Survey of Chain of Thought Reasoning: Advances, Frontiers and Future. 2024. arXiv: 2309.15402 [cs.CL]. url: https://arxiv.org/abs/2309.15402.

[5] Conversica. AI Ethics and Corporate Responsibility Survey Report. Acces-sed: 2024-08-21. 2023. url: https://www.conversica.com/conversation-automation-resources/report/ai-ethics-survey/.

[6] Rina Dechter. "Learning While Searching in Constraint-Satisfaction-Problems." 1986. jan., 178–185. old.

[7] Grégoire Delétang és tsai. Language Modeling Is Compression. 2024. arXiv: 2309.10668 [cs.LG]. url: https://arxiv.org/abs/2309.10668.

[8] J. Deng és tsai. "ImageNet: A Large-Scale Hierarchical Image Database". 2009 IEEE Conference on Computer Vision and Pattern Recognition. 2009, 248–255. old. doi: 10.1109/CVPR.2009.5206848.

[9] Tim Dettmers és tsai. QLoRA: Efficient Finetuning of Quantized LLMs. 2023. arXiv: 2305.14314 [cs.LG]. url: https://arxiv.org/abs/2305.14314.

[10] Voicu D. Dragomir. "The Impact of Intangible Capital on Firm Profitabi-lity in the Technology and Healthcare Sectors". International Journal of Financial Studies 12.1 (2024). issn: 2227-7072. doi: 10.3390/ijfs12010005. url: https://www.mdpi.com/2227-7072/12/1/5.

[11] Shahul Es és tsai. RAGAS: Automated Evaluation of Retrieval Augmented Generation. 2023. arXiv: 2309.15217 [cs.CL]. url: https://arxiv.org/abs/2309.15217.

[12] Alexander L. Fradkov. "Early History of Machine Learning". IFAC-PapersOnLine 53.2 (2020). 21st IFAC World Congress, 1385–1390. old. issn: 2405-8963. doi: https://doi.org/10.1016/j.ifacol.2020.12.1888. url: https://www.sciencedirect.com/science/article/pii/S2405896320325027.

[13] Google. Trend analysis for Deep Learning. Accessed: 2023-11-02. 2023. url: https://trends.google.com/trends/explore?date=all&q=%2Fm% 2F0h1fn8h&hl=en.

[14] Google. Trend analysis for Deep Learning. Accessed: 2023-11-02. 2023. url: https://trends.google.com/trends/explore?date=2019-01-01%202023-10-27&q=generative%20AI&hl=en.

[15] Lane Greene. Language: Finding a Voice. Computers have got much better at translation, voice recognition and speech synthesis. Technology Quarterly. 2017. máj. url: https://www.economist.com/technology-quarterly/2017-05-01/language (elérés dátuma 2023. 11. 02.).

[16] Gauthier Guinet és tsai. Automated Evaluation of Retrieval-Augmented Language Models with Task-Specific Exam Generation. 2024. arXiv: 2405. 13622 [cs.CL]. url: https://arxiv.org/abs/2405.13622.

[17] Geoff Hinton. Neural Networks for Language and Understanding. Creative Destruction Lab Machine Learning. 2017. url: https://www.youtube.com/watch?v=o8otywnWwKc.

[18] Jonathan Ho, Ajay Jain és Pieter Abbeel. "Denoising Diffusion Probabilistic Models". (2020). arXiv: 2006.11239 [cs.LG].

[19] Jeremy Howard és Sebastian Ruder. Universal Language Model Fine-tuning for Text Classification. 2018. arXiv: 1801 . 06146 [cs.CL]. url: https://arxiv.org/abs/1801.06146.

[20] Edward J. Hu és tsai. LoRA: Low-Rank Adaptation of Large Language Models. 2021. arXiv: 2106.09685 [cs.CL]. url: https://arxiv.org/abs/ 2106.09685.

[21] Lei Huang és tsai. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. 2023. arXiv: 2311. 05232 [cs.CL].

[22] Yizheng Huang és Jimmy Huang. A Survey on Retrieval-Augmented Text Generation for Large Language Models. 2024. arXiv: 2404.10981 [cs.IR]. url: https://arxiv.org/abs/2404.10981.

[23] Damjan Kalajdzievski. Scaling Laws for Forgetting When Fine-Tuning Large Language Models. 2024. arXiv: 2401.05605 [cs.CL]. url: https: //arxiv.org/abs/2401.05605.

[24] Alex Krizhevsky, Ilya Sutskever és Geoffrey E Hinton. "ImageNet Classification with Deep Convolutional Neural Networks". Advances in Neural Information Processing Systems 25. 2012, 1097–1105. old. url: https : / / papers . nips . cc / paper _ files / paper / 2012 / hash / c399862d3b9d6b76c8436e924a68c45b-Abstract.html.

[25] Chaeho Chase Lee, Hohyun Kim és Erdal Atukeren. "The Impact of Knowledge Capital and Organization Capital on Stock Performance during Economic Crises: The Moderating Role of a Generalist CEO". Journal of Risk and Financial Management 17.5 (2024). issn: 1911-8074. url: https://www.mdpi.com/1911-8074/17/5/192. https://doi.org/10.3390/jrfm17050192

[26] Szabados Levente. "Az AI eszközök hatása a programozók munkájára és munkerőpiaci kilátásaira". Megjelenés alatt! (2024. június). Megjelenés alatt.

[27] Patrick Lewis és tsai. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. 2021. arXiv: 2005.11401 [cs.CL]. url: https:// arxiv.org/abs/2005.11401.

[28] J. McCarthy és tsai. A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence. https://www-formal.stanford.edu/jmc/ history/dartmouth/dartmouth.html. Accessed: 2024-03-24. 1955. aug.

[29] Charlie Tatenda Mukaro, Abraham Deka és Sylvester Rukani. "The influ-ence of intellectual capital on organizational performance". Future Business Journal 9.1 (2023), 31. old. issn: 2314-7210. doi: 10.1186/s43093- 023-00208-1. url: https://doi.org/10.1186/s43093-023-00208-1.

[30] Long Ouyang és tsai. Training language models to follow instructions with human feedback. 2022. arXiv: 2203.02155 [cs.CL]. url: https://arxiv.org/ abs/2203.02155.

[31] Oded Ovadia és tsai. Fine-Tuning or Retrieval? Comparing Knowledge Injection in LLMs. 2024. arXiv: 2312.05934 [cs.AI]. url: https://arxiv. org/abs/2312.05934.

[32] L. Y. Pratt. "Discriminability-Based Transfer between Neural Net-works". Advances in Neural Information Processing Systems. Szerk. S. Hanson, J. Cowan és C. Giles. 5. köt. Morgan-Kaufmann, 1992. url: https : / / proceedings . neurips . cc / paper _ files / paper / 1992 / file / 67e103b0761e60683e83c559be18d40c-Paper.pdf.

[33] Selvan Sunitha Ravi és tsai. Lynx: An Open Source Hallucination Evaluation Model. 2024. arXiv: 2407.08488 [cs.AI]. url: https://arxiv.org/abs/ 2407.08488.

[34] Dongyu Ru és tsai. RAGChecker: A Fine-grained Framework for Diagnosing Retrieval-Augmented Generation. 2024. arXiv: 2408.08067 [cs.CL]. url: https://arxiv.org/abs/2408.08067.

[35] Jon Saad-Falcon és tsai. ARES: An Automated Evaluation Framework for Retrieval-Augmented Generation Systems. 2024. arXiv: 2311.09476 [cs.CL]. url: https://arxiv.org/abs/2311.09476.

[36] Pranab Sahoo és tsai. A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications. 2024. arXiv: 2402.07927 [cs.AI]. url: https://arxiv.org/abs/2402.07927.

[37] Laura Sevilla-Lara és Erik G Learned-Miller. "Robustness and Plasticity of Object Categorization". arXiv preprint arXiv:2202.05924 (2022). url: https://arxiv.org/pdf/2202.05924.pdf.

[38] Stack Overflow. 2023 Stack Overflow Developer Survey. https://survey. stackoverflow.co/2023. Accessed: 2024-03-24. 2023.

[39] Stack Overflow. 2024 Stack Overflow Developer Survey. https://survey. stackoverflow.co/2024. Accessed: 2024-08-19. 2024.

[40] Keyon Vafa, Ashesh Rambachan és Sendhil Mullainathan. Do Large Lan-guage Models Perform the Way People Expect? Measuring the Human Generalization Function. 2024. arXiv: 2406.01382 [cs.CL]. url: https: //arxiv.org/abs/2406.01382.

[41] Jason Wei és tsai. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. 2023. arXiv: 2201.11903 [cs.CL]. url: https://arxiv. org/abs/2201.11903.

[42] Danna Zheng, Mirella Lapata és Jeff Z. Pan. Large Language Models as Reliable Knowledge Bases? 2024. arXiv: 2407. 13578 [cs.CL]. url: https://arxiv.org/abs/2407.13578.

[43] Kunlun Zhu és tsai. RAGEval: Scenario Specific RAG Evaluation Dataset Generation Framework. 2024. arXiv: 2408. 01262 [cs.CL]. url: https : //arxiv.org/abs/2408.01262.